Population analysis of neural data – developments in statistical methods and related computational models

by

Ziqiang Wei

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

December, 2016

Abstract

A key goal of neuroscience is to understand how the remarkable computational abilities of our brain emerge as a result of interconnected neuronal populations. Recently, advances in technologies for recording neural activity have increased the number of simultaneously recorded neurons by orders of magnitude, and these technologies are becoming more widely adopted. At the same time, massive increases in computational power and improved algorithms have enabled advanced statistical analyses of neural population activity and promoted our understanding of population coding. Nevertheless, there are many unanswered emerging questions, when it comes to analyzing and interpreting neural recordings.

There are two major parts to this study. First, we consider an issue of increasing importance: that many *in vivo* recordings are now made by calcium-dependent fluorescent imaging, which only indirectly reports neural activity. We compare measurements of extracellular single units with fluorescence changes extracted from single neurons (often used as a proxy for spike rates), both recorded from cortical neural populations of behaving mice. We perform identical analyses at the single cell level

and population level, and compare the results, uncovering a number of differences, or biases. We propose a phenomenological model to transform spike trains into synthetic imaging data and test whether the transformation explains the biases found. We discover that the slow temporal dynamics of calcium imaging obscure rapid changes in neuronal selectivity and disperse dynamic features in time. As a result, spike rate modulation that is locked to temporally localized events can appear as a more sequence-like pattern of activity in the imaging data. In addition, calcium imaging is more sensitive to increases rather than decreases in spike rate, leading to biased estimates of neural selectivity. These biases need to be considered when interpreting calcium imaging data.

The second part of this work embarks on a challenging yet fruitful study of latent variable analysis of simultaneously recorded neural activity in a decision-making task. To connect the neural dynamics in different stages of a decision-making task, we developed a time-varying latent dynamics system model that uncovers neural dynamics shared by neurons in a local decision-making circuit. The shared neural activity supports the dynamics of choice generation and memory in a fashion akin to drift diffusion models, and robustly maintains a decision signal in the post-decision period. Importantly, we find that error trials follow similar dynamics to those of correct trials, but their dynamics are separated in shared neural activity space, proving a more correct early decoding estimation of an animal's success or failure at a given trial. Overall, the shared neural activity dynamics can predict multiple measures of

ABSTRACT

behavioral variability including performance, reaction time, and trial correctness, and therefore are a useful summary of the neural representation. Such an approach can be readily applied to study complex dynamics in other neural systems.

In summary, this dissertation represents an important step towards developing model-based analysis of neuronal dynamics and understanding population codes in large-scale neural data.

Primary Reader: Shaul Druckmann

Secondary Reader: Ernst Niebur

Acknowledgments

This work would not have been possible without my advisor, Shaul Druckmann. I am grateful to him for guiding me through this exciting journey and for his continued support. I will always remember Shaul for his forward thinking, exceptional patience and strong work ethic.

I also would like to specially acknowledge Xiao-Jing Wang, whom I regard as my secondary PI. He has made a great effort to supervise me for the computational modeling in working memory and decision making tasks and I have learned tremendously from Xiao-Jing's modeling expertise and unmatched professional insights.

I am forever thankful for the collaboration and mentorship from Nuo Li, Karel Svoboda, Yinan Wan and Philipp Keller. I have enjoyed working with these bright minds a great deal and a large part of my studies depend both on their tremendous efforts to collect data, and their valuable scientific discussions around it.

This work owes a lot to the individuals in Dr. Svoboda's lab, especially Tsai-Wen Chen, Bei-Jung Lin, Hidehiko Inagaki, Kayvon Daie, and Nuo Li, who put years into collecting the electrophysiological and calcium imaging data on which my modeling

is based. I would like to thank them for generously sharing their data, experimental wisdom, and invaluable discussion and feedback.

I specially thank the lab members across all computational labs in Janelia, Tao Hu (Chklovskii lab), Jonathan Amazon, Eizaburo Doi, Kayvon Daie (Druckmann lab), Herve Rouault, Lorenzo Fontolan, Sandro Romani (Romani lab), Jinyao Yan, Srini Turaga (Turaga lab), Ann Hermundstad, and Vijay Samalam. Their input has been very helpful for my scientific development.

I am grateful for my committee members, Ernst Niebur, Louis Scheffer, and Dan O'Connor, for their scientific advice and support during my graduate years, and my former committee chair, Steve Hsiao, and my former advisor, Dmitri Chklovskii, for their generous support and guidance during my early graduate years.

I am thankful for the faculty members and people behind the programs, Solomon H. Snyder Department of Neuroscience at Johns Hopkins University and Janelia Research Campus, including but not limited to Rita Ragan, Beth Wood-Roig, Ulrike Heberlein, Maryrose Franko, Erik Snapp, and Ashley Munteanu, who have generously provided professional advice and kind help at times needed.

I also want to acknowledge Rudiger von der Heydt and Ernst Niebur and the people of their lab whom I have, joyfully and productively, collaborated with for a good period of time during my first year rotation.

I am grateful for my classmates, Thuzar Thein, Shuohao Sun, and Jingjing Sherry Wu for the generous help during my first year of study and lectures.

ACKNOWLEDGMENTS

I am very thankful for my sibling, Cui Tan, for her love and moral support, and caring for my parents during my PhD studies.

I would like to also acknowledge my other friends, Yun Ding, Yuan Chen, Chen Wang, Zengcai Guo, Xi Long, Yingxue Wang, and Wulan Deng, whom I shared my ups and downs with during my years at Janelia.

I dedicate this dissertation to the memory of my mother, Shujie Tan, who has sacrificed so much to get me where I am today. I thank her for doing everything she could to ensure that I received the best education, and for encouraging me to pursue my own dreams. I am forever grateful for her unconditional love and support.

Dedication

This thesis is dedicated to my mother.

Contents

Abstract **ii**

Acknowledgments **v**

List of Tables **xiv**

List of Figures **xv**

1 Introduction **1**

1.1 Development of technologies in large scale recordings of neural data . 7

1.1.1 Multi-electrode array recording 9

1.1.2 Calcium imaging 12

1.2 Analyses of large-scale neural recordings 15

1.2.1 Analyses based on combined datasets 18

1.2.2 Analyses specific for large-scale simultaneous recordings 21

1.2.3 Challenges in analyses of large-scale neural data 23

1.3 Overview of thesis 25

2 A direct comparison of neural dynamics measured with extracellular recording and calcium imaging 29

2.1 Introduction 30

2.2 Results 34

2.2.1 A 'spike to calcium' (S2C) model 35

2.2.2 Hidden dynamics in calcium imaging 40

2.2.3 Biased selectivity in calcium imaging 42

2.2.4 Distortion of network dynamics in calcium imaging 46

2.2.5 Explained variance of temporal dynamics and trial-type selectivity in leading principal components 47

2.2.6 Calcium indicator dynamics can enhance instantaneous decodability of trial-type variables, but delay the observed response to changes in trial epoch 51

2.3 Discussion 54

2.3.1 Biases in imaging dynamics 55

2.3.2 Quantifying the specific variance introduced by indirect measurements of activity 57

2.4 Materials and Methods 58

2.4.1 Electrophysiological and imaging datasets 58

2.4.2 Simultaneous electrophysiology-imaging recordings 59

2.4.3 Description of the 'spike to calcium' model 59

2.4.4 Single neuron analyses 61
2.4.5 Principal component analysis 62
2.4.6 Population decodability analysis of trial type and behavioral epoch 63
2.5 Supplementary 64
2.5.1 Description of calcium-to-spike models 64
2.5.2 Supplementary figure legends 67
2.5.3 Supplementary tables 83

3 Single-trial dynamics of premotor cortex predicts behavioral variability 86

3.1 Introduction 87
3.2 Results 89
3.2.1 Neural dynamics and the shared-activity space 89
3.2.2 Decoders operating on the shared activity space predict behavioral variability 93
3.2.3 Strong switches of neural dynamics at response in ALM 96
3.2.4 Continuous trial identity signals in the shared activity space . 100
3.2.5 Error trials 101
3.3 Discussion 105
3.4 Materials and Methods 108
3.4.1 Electrophysiological recordings 108

3.4.2 Single neuron analysis 109
3.4.3 Coding directions and neural dynamics in coding directions . . 110
3.4.4 Time-varying linear dynamics systems model and neural mode dynamics 111
3.4.5 Leave-one-neuron-out estimation and optimal dimension of the shared-input space 113
3.4.6 Reaction time correlations 115
3.5 Supplementary 116
3.5.1 Supplementary figures 116
3.5.2 Supplementary tables 123

4 Confidence Estimation as a Stochastic Process in a Neural Dynamical System of Decision Making 132

4.1 Introduction 133
4.2 Results 136
4.2.1 Network dynamics in a fixed duration task with post decision wagering at zero motion strength 139
4.2.2 Behavioral performance 144
4.2.3 Choice confidence as a logistic function of the differential activity 146
4.2.4 Low confidence results in changes of mind to sure target . . . 156
4.2.5 A sure target as a probe about the system's confidence 160
4.2.6 Assessment of choice confidence in a reaction time task 162

4.3 Discussion 166

4.3.1 Comparison with existing models 169

4.4 Materials and Methods 175

4.4.1 Network model 175

4.4.2 Simulation protocol of fixed-duration discrimination decision task 176

4.4.3 Measurements of activity trajectories 181

4.4.4 Choice confidence assessment 181

5 Conclusions and future directions **184**

A General Methods in population analysis in neural data **192**

A.1 Sparse linear discriminant analysis 193

A.1.1 Description of linear discriminant analysis 196

A.1.2 Sparse linear discriminant analysis 198

A.2 Time-varying linear dynamical system analysis 200

A.2.1 Description of linear dynamical system analysis 202

A.2.2 Leave one neuron out estimation 209

A.2.3 Time-varying latent dynamical system analysis 212

A.2.4 Performance of time-varying latent dynamical system analysis on neural data 213

Bibliography **223**

List of Tables

2.S1 Delayed discrimination task 84
2.S2 Simultaneous ephys-imaging experiments 84
2.S3 Spike-to-calcium model parameter range 84
2.S4 Spike-to-calcium model parameter sensitivity 85

3.S1 Correct trial information 126
3.S2 Error trial information 127
3.S3 Recording methods, depth and cell type information 128
3.S4 Explained variance 129
3.S5 Neural dynamics and reaction time correlation 130
3.S6 Effective eigenvalues 131

List of Figures

2.1 Simultaneous recordings of ALM population using different techniques 34
2.2 Schematic description of spike-to-calcium model 37
2.3 Neuronal selectivity measured by calcium imaging exhibits less heterogeneity in temporal dynamics than when measured by electrophysiology 40
2.4 Induction of distinct activity dependent biases in different populations of neurons by calcium dynamics 44
2.5 Calcium imaging exhibits more sequence-like population activity than that of ephys recordings 45
2.6 Temporal dynamics account for most variance in first principal component of ephys data, but trial type selectivity accounts for most variance in calcium data 48
2.7 Calcium imaging data show a delayed increase of selectivity 50
2.S1 Supporting figure for Figure 2.2: details in spike-to-calcium models . 68
2.S2 Supporting figure for Figure 2.3: dynamics of selectivity in different imaging conditions 71
2.S3 Supporting figure for Figures 2.2-2.3: details in calcium-to-spike (C2S) models and inferred ephys from imaging using C2S model can account a fraction of multiphasic neuron in imaging 73
2.S4 Supporting figure for Figure 2.5: inferred ephys from imaging using calcium-to-spike model can account for little time-lock dynamics . . . 78
2.S5 Supporting figure for Figure 2.6: explained variance of first three PCs are robust to the specificity of the confounding factors in comparison 81

3.1 Neural activity of anterolateral motor cortex neurons in shared-activity space 91
3.2 Trial identity decoded from TLDS model correlates with trial-by-trial variability in behavioral reaction time and performance on single trials 95
3.3 Neural activity of anterolateral motor cortex neurons exhibits a discontinuity of selectivity at response 96
3.4 Two hypothesized models of trial identity dynamics 97

3.5 Maintenance of trial identity across dynamical transition revealed by time-varying linear dynamical systems model 98
3.6 TLDS model still has predictive power in error trials and reveals failure of trial identity maintenance in error trials 103
3.S1 Neural activity in ALM exhibits a switch of selectivity at response in the 1^{st} PCA space 118
3.S2 Fitting details of time-varying linear dynamical system models 119
3.S3 Rank correlation of trials is higher in shared-input space than that in boxcar-smoothed full-neural space 122
3.S4 Rank correlation of trials using neural dynamics in Gaussian Process Factor Analysis space 124
3.S5 Rank correlation between neural-mode LDA score and reaction time . 125

4.1 Schematic description of the decision task and model architecture . . 138
4.2 Neuronal activity of sample trials at zero motion strength 141
4.3 Behavioral performance 146
4.4 Differential activity of two competing choices determines whether a sure target is waived 148
4.5 Onset time of sure target determines the probability of choosing sure target but has little impact on accuracy 150
4.6 The probability of waiving T_s reflects choice confidence 152
4.7 Low confidence results in changes of mind to sure target in post-decision wagering 158
4.8 Effect of sure target input strength on the behavioral performance . . 160
4.9 Choice confidence in a reaction time task 164

A1.1 Schematic description of linear discriminant analysis for neural code of trial type 194
A1.2 Schematic description of linear dynamical system analysis and leave-one-neuron-out estimation 208
A1.3 Performance of time-varying latent dynamical system analysis on the artificial data using Gaussian variability 215
A1.4 Performance of time-varying latent dynamical system analysis on the artificial data using Poisson variability 218
A1.5 Performance of time-varying latent dynamical system analysis on the real simultaneous neural recordings 221

Chapter 1

Introduction

A key question in systems neuroscience is how to describe the properties and underlying principles of operation of large-scale complex neural networks, and in particular considering the fact that our examination of them is typically based on data that are severely incomplete in many ways (Dayan and Abbott, 2001). Even a simple cognitive operation often involves the cooperation and coordination of millions of neurons, but we are only able to access the activity of a very small fraction of this population, i.e. at most a few hundred neurons at a time and with little explicit knowledge of important properties such as their detailed biophysics or the structure of their connectivity. This challenge sets the importance of advanced data acquisition and analysis approaches comprising firstly, the simultaneous neural recordings, secondly, state-of-the-art statistical models of data mining, and lastly, the use of computational modeling, to establish a foundation upon which we can develop our

understanding of neural circuitry at a mechanistic level and generate new predictions for designing new experiments.

For many years, most recordings of neural activity were restricted to a handful of neurons at a time (which also lacked chronic stability), limiting one's interpretation to properties and functions of single neurons instead of local neural circuits. Of necessity, experimenters would commonly determine the receptive fields or the response fields of single neurons, instead of covariance across a neural population. Accordingly, there was only limited need for statistical models of neural data population analysis. At the time, computational scientists tended to focus on simplified models of ideal neural circuits. Though very useful, these models tended to be highly abstract, and were difficult to relate to the detailed properties of neuronal dynamics and structure.

Arguably, understanding brain function requires monitoring and interpreting the activity of large neuronal networks during specific behaviors. Recent development of modern recording technologies, based on both multi-electrode arrays and optical imaging, has enabled our ability to simultaneously access hundreds or even thousands of neurons (Briggman et al., 2005; Ahrens et al., 2012; Ahrens and Engert, 2015; Vladimirov et al., 2014; Keller and Ahrens, 2015; Ahrens et al., 2013). This helped pave the way to a new era of neuroscience. Nevertheless, given the growing size and complexity of neural recordings, analyzing and interpreting the data willlikely pose a fundamental bottleneck for neuroscience (Stevenson and Kording, 2011). For example, an hour of two-photon imaging in mice can yield hundreds of gigabytes of

spatiotemporal data, and recordings from nearly the entire brain of a larval zebrafish using light-sheet microscopy can yield several terabytes comprising the activity of more than a hunderd thousand neurons. At this rate, novel statistical models for population analysis are unprecedentedly in demand to (1) process the neural data at high speeds, (2) and to interpret the meaningful population's dynamics in high dimensional neural space (Cunningham and Yu, 2014; Harris et al., 2016; Freeman et al., 2014; Freeman, 2015; Ji et al., 2016).

In addition, current experiments are performed in situations in which the detailed knowledge of the connectivity structure among the recorded neurons is largely unknown. In principle, one can infer it from data using computational models (Lim et al., 2015; Engel et al., 2015) and determine to which degree a connectivity pattern would explain the dynamics observed. This is often done in a model-based way; computational scientists base their model assumptions on comprehensive statistics of the neurophysiological data. More importantly, they must be willing to engage in the never-ending cycle of modifying theories given new insights from the data, updating model predictions, and revisiting the data to test their predictions, until a more accurate picture of the subject matter appears.

This dissertation is dedicated to the use of large-scale neural recordings and a data-driven modeling approach to study one of the most fundamental cognitive processes in animals: decision making. The brain has evolved intricate deductive machinery to emancipate us from the simple immediate and reflexive response, and substitute more

flexible decision-making skills (Gold and Shadlen, 2007; Shadlen and Kiani, 2013). This capability empowers us to process sensory data and guides appropriate behavioral responses. Moreover, the neural circuits involved in these processes may also be the building blocks of more sophisticated aspects of human cognition. For example, brain circuits support integrating evidence from diverse sources (e.g. different sensations and memory), assigning levels of importance to cues that differ in reliability, calculating expected costs and benefits associated with anticipated outcomes (Yang and Shadlen, 2007), and holding the decision in memory until an action is required (Shadlen and Newsome, 2001; Roitman and Shadlen, 2002). This is a multi-step process for making sense of the external world and selecting appropriate actions in different situations, which is vital for survival, where determining whether another being is predator or prey, whether a food item is poisonous or nutritious, or whether a situation is dangerous or safe can mean life or death. Overall, decision making serves as a window into cognition, for which analysis of neural dynamics is of particular interest.

Decision making as a high-order cognitive behavior is thought to be mainly supported by neuronal circuits in the frontal and parietal cortices (Gold and Shadlen, 2007; Shadlen and Newsome, 2001; Shadlen and Shohamy, 2016; Shadlen and Kiani, 2013; Kepecs et al., 2008; Romo et al., 1999; Guo et al., 2014b; Guo et al., 2015; Brody et al., 2003; Stuphorn and Schall, 2006; Stuphorn et al., 2010). Since single neurons in these areas exhibit highly variable activity even in seemingly identical

trials (Brody et al., 2003; Romo et al., 1999; Rigotti et al., 2013), one would expect to uncover the neural signal of decision making more reliably at the population level. The population description of such neural signals has become available with the application of simultaneous recording technologies (multi-electrode arrays and optical imaging) (Li et al., 2015). Nevertheless, central issues regarding the extraction of neural signals from high-dimensional time-series recording data remain largely unknown. Here we focus on two such questions: one relates to the data recording approach itself, and the other relates to the extraction of neural dynamics in different states of decision making (e.g. choice generation, memory and response).

Decision-making neural circuits have been long investigated using electrophysiological recordings (Gold and Shadlen, 2007; Shadlen and Kiani, 2013). With the development of highly sensitive fluorescent protein-based indicators and powerful new imaging methods, calcium imaging has been widely adopted for measurements of neural population activity (Tian et al., 2009; Chen et al., 2013; Akerboom et al., 2012; Pologruto et al., 2004; Ohkura et al., 2012; Inoue et al., 2015; Dana et al., 2016). The fidelity of recording is an important question in the scientific research of population activity, since calcium imaging only indirectly reports spiking activity. Comparing recordings performed with electrophysiology or imaging under indentical behaviors, we find that the neural dynamics of decisions show several differences that could arise from distinct recording technologies. The transformation from spikes to calcium is non-linear due to the dynamics of intracellular calcium concentration and

nonlinearities imposed by the use of protein-based indicators (Scheuss et al., 2006; Tian et al., 2009; Chen et al., 2013; Akerboom et al., 2012; Pologruto et al., 2004). One has thus to consider this transformation based on the spike-to-fluorescence mechanism, before the interpreting dynamics directly from calcium imaging recordings. The first line of work in this dissertation concerns the spike-to-fluorescence mechanism and the characterization of the differences between dynamics measured by electrophysiology and imaging.

The second part of this work addresses statistical models designed to uncover low-dimensional dynamics from high-dimensional neuronal activity time series obtained under multi-state behavioral conditions. Because the world is not static, we often need to base our response upon immediate sensory inputs, which may be connected to recent or distant memories and experiences. Decision making could thereby consist of multiple stages, associated with behavioral states, especially in the well-established experimental paradigm of the delayed discrimination task. In different states of the behavior, the neural dynamics may adapt to follow the specific properties of computations that underlie the different phases of choice generation and memory. More importantly, these principles of computation in different phases of the decision have not been examined at the population level. For example, both the drift diffusion model (Ratcliff and Smith, 2004; Ratcliff and Starns, 2009; Kiani et al., 2008; Kiani and Shadlen, 2009; Kiani et al., 2014; Mazurek et al., 2003) and the attractor model (Furman and Wang, 2008; Wang, 2002; Wang, 2008; Wei and Wang, 2015;

Wong and Wang, 2006; Wong et al., 2007) can explain the neural dynamics of single neurons in decision making tasks, however, they have distinct predictions at the population level (for example, the drift diffusion model assumes that the sensory input contributes equally across time to the final decision, while the attractor model emphasizes that early sensory input plays a dominant role towards decision). Little is known about the explanatory power of computational models in large-scale neural recordings. This dissertation is one of the early attempts to shed light on this issue.

The rest of the introduction provides a brief background concerning simultaneous population recording technologies and related developments in data analysis suitable for general readers. We will also guide readers through the two questions of interest in analyses of large-scale simultaneous recordings mentioned above. Advanced readers may refer to the introduction sections within each chapter for summaries of our main results. Readers with selective interest can also refer to the overview of all topics in this dissertation at the end of this chapter (**Section 1.3**).

1.1 Development of technologies in large scale recordings of neural data

Electrical activity is an intrinsic property of a neuron whose links to the function of nervous systems and animal behavior has long sparked the imagination of scientists (Galvani and Aldini, 1792). For generations, neuroscientists have continually

developed and advanced electrophysiological tools that allow us to probe all levels of neural activities, from the dynamics of a single ion channel to the spiking activity of hundreds of neurons in a local network. In particular, an electrode is the tool often used to directly monitor spiking events, a basic currency for communication among neurons. It monitors the electrical activity of neurons adjacent to the electrode tip, the size of which determines the number of neurons being monitored simultaneously. Furthermore, multi-electrode arrays are employed to span even larger spatial ranges for simultaneous recordings. In electrophysiology, noise comes primarily from recording instruments, and the signal-to-noise ratio has been maximized to allow resolution of the opening of single ion channels. The direct recording of electrical activity, like spike events, with high signal-to-noise ratios is thus the main strength of the method (Scanziani and Hausser, 2009). However, because of the instability of its mechanical implementation, the electrode is difficult to use for long-term monitoring of single cell activity (Nicolelis et al., 2003).

Optical imaging provides an alternative way to probe neural activity, by observing the dynamics of an indicator (reporter), e.g. voltage sensors for recording of membrane voltage changes and calcium sensors for recording of calcium concentration. Such an indicator typically takes the form of a molecule that converts membrane potential (or its consequences) into a more easily observed optical signal. The optical imaging is thereby an indirect way of recording neural activity. Optical imaging offers several key advantages: (1) exceptional spatial resolution that allows reporting of sig-

nals in small neuronal structures, like dendritic spines (Ji et al., 2008; Sun et al., 2015; Wang et al., 2015; Hell, 2007; Hell, 2010; Wilt et al., 2009); (2) the possibility for simultaneous recordings across a large spatial scale, even the whole brain (Ahrens et al., 2012; Ahrens and Engert, 2015; Vladimirov et al., 2014; Keller and Ahrens, 2015; Ahrens et al., 2013); (3) targeting specific cellular subtypes and sub-cellular domains, when accompanied by genetic tools (Luo et al., 2008); (4) chronic recording of the same group of neurons, even throughout an animal's lifetime (Dana et al., 2014). There is however a downside to indirect reporting. The properties of indicators and optical detection systems (like those of the microscope and imaging camera) can limit the temporal resolution or signal-to-noise of the recordings.

To demonstrate the difference between direct and indirect simultaneous recordings, we will compare multi-array electrophysiological recordings with calcium imaging from experiments, where both were recorded simultaneously, and make more quantitative comparisons at the level of neural dynamics in behavior relevant conditions (**Chapter 2**). We will also show to which degree the difference of neural dynamics in both recording conditions can be explained by the dynamics of the calcium indicator (discussed in **Chapter 2**).

1.1.1 Multi-electrode array recording

The spike is the fundamental currency in neuronal communication. Electrophysiological recordings can be used to infer spikes in a fairly straightforward fashion (in

most cases) with high temporal fidelity and signal-to-noise ratio, making it the 'gold standard' for neuronal signaling study (Scanziani and Hausser, 2009). The signal-to-noise ratio is a fundamental consideration when comparing electrophysiology and optical imaging. However, the spatial resolution and simultaneous sampling size of an individual electrode is limited by its tip size in the recordings. To improve the sampling in space, large numbers of electrodes with fine spacing have been assembled as multi-electrode arrays in recordings of population activity. Technically, a multi-electrode array is a device that includes multiple plates or shanks through which neural signals are sampled. Two general classes of multi-electrode arrays are applied in different recording conditions: one is implantable, used *in vivo*, and the other is non-implantable, used *in vitro*.

In the 1950s, the multi-electrode array was first applied in simultaneous recordings with a handful of units (Cheung, 2007; Nicolelis, 2007). This was followed by tremendous growth in the number of simultaneously recorded units, and that number has doubled approximately every seven years, over the last five decades (Stevenson and Kording, 2011; Spira and Hai, 2013). Currently, an *in vitro* multi-electrode array may contain over 10,000 electrodes (Hutzler et al., 2006; Berdondini et al., 2009; Frey et al., 2009; Nam and Wheeler, 2011; Huys et al., 2012) and an *in vivo* array may have over a hundred (Hochberg et al., 2006; Schwartz, 2004). Following the developing trend of electrode technology, one would expect to be recording from thousands of neurons in the next two decades. However, the chronic stability, e.g.

tissue displacement and invasive contact with the local neural system, may fundamentally limit the density with which electrodes can be implanted (Nicolelis et al., 2003). Moreover, on the computational side, the efficiency of spike sorting may also be a substantial bottleneck for large-scale multi-electrode recordings (Brown et al., 2005; Stevenson and Kording, 2011).

Multi-electrode recording is still a better solution for deep-layer recordings, compared to optical methods, since the recording depth is constrained in optical imaging. Light scatters and is attenuated as it passes through tissue. This reduces the intensity of the signal that can be sampled from deeper brain areas, and is a hard technical barrier that is difficult to overcome in optical imaging. For instance, high-resolution one-photon imaging has been limited to thin preparations, or only the most superficial regions (depths $< 50\mu m$) of intact tissue. Two-photon imaging using nonlinear microscopy is currently still limited to the superficial regions of the brain ($< 1,000\mu m$). A recent development in micro-endoscopy inserts a probe into the brain region of interest, however this is typically considered to be a more invasive method of neural recording than classical electrophysiological approaches (Wilt et al., 2009).

Electrophysiology is still an evolving discipline at present, and its machinery is being refined for the new applications. Notably, there are a series of developments underway that aim to overcome some of the key traditional limitations of electrophysiology. First, using nano-fabrication techniques, ever-smaller electrodes (with tip size $< 1\mu m$) are being produced to record neural activity from extremely

fine structures such as boutons and spines (Qiao et al., 2005; Krapf et al., 2006; Heller et al., 2005). Second, multi-electrode arrays with larger numbers of electrodes are being assembled with ever-finer spacing to improve spatial sampling in recordings of population activity (Miller and Wilson, 2008; Smith et al., 2004). Finally, research into the physical basis for the long-term interaction between electrodes and neural tissue could eventually allow invasive brain-machine interfaces to target populations of neurons more precisely, more reliably, and over longer timescales (Fromherz, 2006).

1.1.2 Calcium imaging

Optical imaging is an indirect readout of neural activity through voltage or calcium sensors, in which the final neural recordings rely both on the sensitivity of the indicators and on the detectability of their signals by the imaging system. For example, the electrical signals in neurons can be as fast as < 1 ms, while the kinetics of an indicator could be far slower. This places a severe constraint on the biochemical kinetics of the indicators and the detection system, where fast responding indicators and rapid scanning technologies are required in neural recordings. The current recordings made by optical imaging suffer from low temporal resolution and signal-to-noise ratio, both from the properties of the indicators and optical detection systems. Nevertheless, optical imaging has already surpassed electrophysiology where high spatial resolution and genetic specificity are required when measuring neural activity.

Why are calcium sensors of particular interest for optical imaging monitoring of

neural dynamics? Calcium is a major signaling molecule in neurons, and synaptic inputs, and membrane voltage fluctuations often trigger changes in intracellular calcium concentration. Hence, calcium indicators have long been successfully used to infer both sub- and supra-threshold activity in neurons (Berger et al., 2007). Since organic calcium-sensitive dyes are sensitive enough to respond to the opening of a single calcium-permeable channel in a spine, when detected by two-photon microscopy, these dyes can be used to monitor the occurrence of both action potentials and synaptic input to spines (Nimchinsky et al., 2002; Palmer and Stuart, 2009; Sabatini and Svoboda, 2000; Yuste et al., 1999; Matsuzaki et al., 2001; Mizrahi et al., 2004; Sabatini et al., 2002; Xu et al., 2012).

Although scattering of light, brain movement, and the unknown dendritic distribution of active synaptic inputs have so far limited the direct detection of synaptic input patterns, membrane-permeant calcium dyes have been used to successfully monitor network activity in neurons and glia cells within the intact brain, under both anesthetized and awake conditions (Ohki et al., 2005; Stosiek et al., 2003; Greenberg et al., 2008; Dombeck et al., 2007; Wang et al., 2015). However, even with the best existing indicators, the prolonged time course of the intracellular calcium signal triggered by action potentials, coupled with the limitations of *in vivo* imaging, have made it challenging to reliably detect single action potentials in behaviorally relevant conditions.

The design of genetically encoded calcium indicators has attracted intense in-

terest in recent years. These indicators are typically based on a calcium-sensitive molecule, such as calmodulin or troponin, fused to GFP or other fluorescent proteins, with calcium binding reported by fluorescence changes due to alterations in the efficiency of fluorescence resonance energy transfer or changes to the chromophore environment. Several generations of sensor development have yielded vastly improved properties, particularly a family of ultra-sensitive genetically encoded calcium sensors (the GCaMP families) that outperform other sensors in terms of brightness, temporal resolution and signal-to-noise ratio, in cultured neurons and in zebrafish, flies and mice *in vivo* (Chen et al., 2013). Furthermore, the genetically encoded calcium indicators can be applied to transgenic animals. These animals are capable of stably reproducing the specific calcium indicators under the control of a promoter (*Thy1* for instance), which lends itself to a life-term cellular imaging of neuronal populations in the intact brain (Dana et al., 2014).

Finally, to combine the advantages of different recording techniques, there is a modern trend to integrate the use of electrophysiology and calcium imaging, e.g. the combination of optical imaging and single-unit or multi-unit recordings to reveal the role of single neurons in network dynamics (Arieli et al., 1996; Tsodyks et al., 1999; Katzner et al., 2009), and the combination of patch-clamp recording and two-photon population imaging to map functional connectivity in networks. Such integration of recording techniques should significantly quicken the pace of discovery as we move towards the goal of linking brain structure to their function within the scope of

neuroscience.

1.2 Analyses of large-scale neural recordings

A central goal of systems neuroscience is to link the dynamics of neural circuits to behavior. Particularly, large-scale neural recordings have begun to shed light onto cellular-level observations of the function and organization of the nervous systems (Ahrens et al., 2012; Ahrens and Engert, 2015; Vladimirov et al., 2014; Keller and Ahrens, 2015; Ahrens et al., 2013). These advances in neural recordings beg important consideration for emerging data analysis techniques (Stevenson and Kording, 2011; Freeman et al., 2014).

What information would the neural activity represent? This is the central question of neuron coding (Dayan and Abbott, 2001). On the encoding side of the question, one wants to know how information from the external world is encoded in neuronal spikes. On the decoding side of the question, one aims to use neural activity to predict behavior and ultimately apply this knowledge in design of translational applications, such as brain-computer interfaces.

Traditionally, as simultaneous recording was limited to single or a few neurons, neuroscientists determined the information carried by a neuron using its receptive field to external input or its response field to a movement. This is usually done

by using neural responses to a specific behavioral condition modeled on the average across trials (within a given condition) and smoothed into a peri-stimulus time histogram (Dayan and Abbott, 2001). However, a neuron could be a hub in the network that collects multiple stimuli for external inputs, movements, or some other unknown internal brain states; its dynamics could therefore be modulated by multiple behavioral parameters of a task and its receptive field or response field could be mixed. In general, the traditional analysis of a single neuron could yield neural responses with mixed preferences that are difficult to interpret (Mante et al., 2013).

With the advance of simultaneous recording techniques, computational neuroscientists have re-examined the neuronal properties at the population level (Cunningham and Yu, 2014; Peron et al., 2015; Harris et al., 2016; Freeman et al., 2014; Freeman, 2015; Ji et al., 2016). Particularly, one class of statistical methods, i.e. dimensionality reduction, was developed to extract simple structure from these seemingly complex data (Roweis and Ghahramani, 1999). The underlying hypothesis of these statistical methods is simple yet powerful (Cunningham and Yu, 2014). Imagine we are examining the neural dynamics in simultaneous neural recordings for a small local neural circuit. The neuronal responses are often highly correlated with each other, such that one could hypothesize that the recorded neurons belong to a common underlying network and the covarying activity stems from a smaller number of explanatory variables. In modern statistics, dimensionality reduction methods are employed to discover and extract these explanatory variables from the high-dimensional data;

the resulting explanatory variables are often termed latent variables since they are not directly observed, and any data variance not captured by the latent variables is considered to be "noise". In neuroscience, there is a more meaningful way to interpret latent variables, which are often called neural modes. The neural modes can be thought of as common inputs or, more generally, as the collective role of unobserved neurons in the same network as the recorded neurons. In this case, neurons in the common underlying network are viewed as the nodes, the dynamics of which are driven by multiple neural modes and probably some unknown independent input, usually referred to "noise", and vice versa, each neural mode can be modeled simply by a weighted combination of linked neuronal activity.

Since the neural mode represents the dynamics of common drives to the neurons, one can combine the usage of dimensionality reduction methods and time series analysis to characterize the independent stochastic drive, like spiking variability, onto single neurons in time, and across neurons. The goal of such a dimensionality reduction approach is to characterize how the firing rates of different neurons covary and to discard the spiking variability as noise. The neural modes therefore define a low-dimensional space that represents shared neural dynamics that are prominent in the population response. On the other hand, by projecting neural mode dynamics back to the original neural space, dimensionality reduction attempts to find the neural modes that can reconstruct the population activity as well as possible. The reconstructed activity can be interpreted as the de-noised firing rate for each neuron.

In the following two sections, we will introduce two classes of dimensionality reduction methods for practical use in neural data: one is for general situations, where the data are collected non-simultaneously; the other is specific to a simultaneously recorded dataset. Both methods can be used interchangeably, given the appropriate situations. In **Chapter 5**, we will discuss which questions would be better answered using simultaneous recordings, and by specific analysis methods.

1.2.1 Analyses based on combined datasets

Although the advent of multi-electrode recordings dates back to the 1950s, the high cost of hardware and software limited the spread of multi-electrode arrays until the 1990s (Fejtl et al., 2006; Pine, 2006). Analysis of population codes was therefore either based on pairwise correlation across a few neurons (Brillinger, 1992; Gerstein and Perkel, 1969; Abeles, 1982; Kass et al., 2003; Ventura et al., 2005; Cohen and Maunsell, 2009; Cohen and Newsome, 2008) or done on the collection of neuronal activities that were recorded in separate trials. In the latter case, given the lack of simple interpretation at the level of individual neurons, one would ask whether the confounding single-neuron responses could be understood as views of a simple dynamic process at the population level. Dimensionality reduction is a common way to approach this. One can apply the methods either to the "de-noised" neural data, which is averaged across time (or even smoothed over time using some predefined filter), or to a collection of pseudo-simultaneous recording sessions generated from

non-simultaneous recorded trials. We will introduce here two classes of dimensionality reduction methods: unsupervised learning based methods and supervised learning based methods.

Principal component analysis (PCA) (Jolliffe, 2014; Mazor and Laurent, 2005; Churchland et al., 2010a; Freeman et al., 2014) and factor analysis (FA)(Jvreskog, 1996; Knott and Bartholomew, 1999; Churchland et al., 2010b; Sadtler et al., 2014; Santhanam et al., 2009) are two of the most basic and well-used unsupervised learning dimensionality reduction methods, and follow linear models. PCA identifies an ordered set of orthogonal directions that captures the greatest variance in the data. Since the data is considered to be"de-noised" in preprocessing, one would imagine that most of the information is in the first few dimensions that capture the most of the variance. Although capturing the largest amount of variance may be desirable in some scenarios, one caveat is that the low-dimensional space identified by PCA captures variance of all types, including firing rate variability and spiking variability. PCA is usually applied to trial-averaged neural dynamics and is of limited utility when explaining the single-trial dynamics. FA can be used to better separate changes in firing rates from spiking variability. FA identifies a low-dimensional space that preserves variance that is shared across neurons (firing rate variability), while discarding variance that separates each neuron (spiking variability). FA can pass for PCA with the addition of an explicit noise model that allows FA to discard the independent variance of each neuron. In addition

to these linear methods (Roweis and Ghahramani, 1999), nonlinear models have been developed to uncover low-dimensional nonlinear manifolds in high-dimensional space. Two of the most prominent methods to identify nonlinear manifolds are Isomap (Tenenbaum et al., 2000) and locally linear embedding (Broome et al., 2006; Saha et al., 2013; Stopfer et al., 2003; Brown et al., 2005; Carrillo-Reid et al., 2008; Roweis and Saul, 2000). Both nonlinear methods use local neural spaces to estimate the structure of the manifold. Such estimation is sensitive to sampling bias (where the full neural space is not explored) and are fragile in the presence of noise (Boots and Gordon, 2012), limiting the use of nonlinear methods compared to the linear ones.

The other often-used class of statistical methods is based on supervised learning, such as linear discriminant analysis (LDA) (Briggman et al., 2005; Durstewitz et al., 2010; Bartho et al., 2009; Li et al., 2016), and de-mixed principal component analysis (dPCA) (Machens, 2010; Brendel et al., 2011; Kobak et al., 2016). In experiments, we often associate population activity with some simultaneously recorded behavioral variables, e.g. stimulus identity, decision identity, etc. We could label each identity of the behavior variable as a group (Stimulus A vs Stimulus B). A possible objective of dimensionality reduction is to project the data such that differences in these groups are maximized. Specifically, LDA can be used to find such a low-dimensional projection, in which the between-group variance of neural dynamics is maximized relative to the within-group variance. In the case of multiple behavioral parameters (such as

Stimulus × Decision), one can either make a tensor group collection that considers all possible combinations of groups, or seek to "de-mix" the effects of different behavioral parameters, such that each projection of the neural data captures the variance of a single behavioral variable. In principal, these methods can be unified as a variant of generalized linear regression (GLM) (Mante et al., 2013; Machens et al., 2010; Pillow et al., 2008), which incorporates the behavioral variable with a few discrete identity values or a continuum of identity values.

1.2.2 Analyses specific for large-scale simultaneous recordings

With the advent of stable simultaneous recording technology, such as multi-electrode arrays and optical imaging, it has become common to exploit modern statistical models to probe further into population dynamics in single trials (Afshar et al., 2011; Ames et al., 2014; Gao et al., 2015; Gilja et al., 2012; Kao et al., 2015; Kaufman et al., 2014; Kaufman et al., 2015; Macke et al., 2011; Shenoy et al., 2011; Yu et al., 2009; Petreska et al., 2011). By definition, spontaneous activity involves fluctuations of the population activity that are not directly controlled by the experimenter. To characterize spontaneous activity, dimensionality reduction can be applied to extract a low-dimensional network state that incorporates the analysis in time. This facilitates the comparison of spontaneous activity to population activity

during sensation and action (Afshar et al., 2011; Gao et al., 2015; Gilja et al., 2012; Kao et al., 2015; Kaufman et al., 2014; Kaufman et al., 2015; Shenoy et al., 2011; Petreska et al., 2011).

Why is simultaneous recording data particularly needed in the study of single trials? This is because the temporal dynamics provide extra information to examine the independent drive onto single units in the neural data. If the data form a time series, one can leverage the sequential nature of the data to provide further de-noising and to characterize the temporal dynamics of the population activity. Here, we focus on unsupervised analysis, which will correlate with our latent-variable model shown in **Chapter 3** and **Appendix A**. There are several dimensionality reduction methods available for time series: hidden Markov models (HMM) (Seidemann et al., 1996; Jones et al., 2007; Ponce-Alvarez et al., 2012; Bollimunta et al., 2012; Abeles et al., 1995; Danóczy and Hahnloser, 2006; Kemere et al., 2008; Morcos and Harvey, 2016), kernel smoothing followed by a static dimensionality reduction method (Yu et al., 2009), Gaussian process factor analysis (GPFA) (Yu et al., 2009), latent linear dynamical systems (LDS) (Smith and Brown, 2003; Kulkarni and Paninski, 2007; Paninski et al., 2010; Buesing et al., 2012; Pfau et al., 2013) and latent nonlinear dynamical systems (NLDS) (Petreska et al., 2011; Macke et al., 2011). All of these methods return low-dimensional, latent neural trajectories that capture the shared variability across neurons for each high-dimensional time series. An HMM is applied in settings where the population activity is believed to jump between discrete states,

whereas all of the other methods identify smooth changes in firing rates over time, where the degree of smoothness is determined by the data.

Unlike the combined dataset analysis, where one would only obtain the trial-averaged responses across a population of neurons, single-trial population dynamics can be extracted using HMM, GPFA, LDS or NLDS. These methods yield single-trial neural trajectories, which facilitate the comparison of population activity across trials, and a low-dimensional dynamics model, which characterizes how the population activity evolves over time (Cunningham and Yu, 2014). These methods are particularly appropriate for single-trial population activity because they model explicitly the noise within each neuron. As a cautionary note, the dynamics model in GPFA is stationary and encourages smoothing the neural trajectories, while applying LDS or NLDS to fit the data, one realizes that (1) the dynamics model is generally non-stationary and (2) the neural trajectories often follow a set path within the dynamics.

1.2.3 Challenges in analyses of large-scale neural data

Throughout this dissertation, our work is based on the neurophysiological data collected in the laboratory of Dr. Karel Svoboda at Janelia Research Campus, HHMI. For years, Dr. Svoboda and his colleagues have conducted a series of crucial experiments that have become a cornerstone in our understanding of neural

circuits in sensory processing and motor planning (Guo et al., 2014a; Guo et al., 2014b; Huber et al., 2012; Komiyama et al., 2010; Li et al., 2015; Li et al., 2016; O'Connor et al., 2010; O'Connor et al., 2013; Peron et al., 2015; Peron et al., 2015; Sofroniew et al., 2016; Yu et al., 2016; Dana et al., 2014). In a typical paradigm of these experiments (Guo et al., 2014a; Guo et al., 2014b; Li et al., 2015; Li et al., 2016; O'Connor et al., 2010; O'Connor et al., 2013; Peron et al., 2015; Peron et al., 2015; Yu et al., 2016; Dana et al., 2014), mice were trained to perform a delayed discriminant task, where a sample stimulus is presented at different locations relative to the mouse's whiskers, followed by a brief delay. To receive a reward, the mouse must perform one of a set of actions, specifically the action that is associated with that stimulus (e.g., lick left for one stimulus condition, and lick right for another). Neural dynamics were studied by both electrophysiology and calcium imaging to examine the different aspects of the same local circuit. Interestingly, there are some notable differences in neural dynamics when sampled with different recording technologies. Our first collaborative work, in **Chapter 2**, is thus to determine to which degree we can predict and undo such differences using spike-to-calcium dynamics. At the same time, a set of the recordings was obtained using multi-electrode arrays, including several sessions with tens of simultaneously recorded units. In another collaborative study (**Chapter 3**), we examine whether more information could be uncovered in single-trial analysis compared to the traditional analysis based on the combined data. As mentioned above, these two questions present two fundamental challenges in anal-

ysis of large-scale neural data, which are examined in this thesis. One result is our early work to unify the data sampled from different technologies, potentially important since the mixed use of electrophysiology and optical imaging will be a trend for the next few decades. The other is one of the first few studies to uncover the switch of neural dynamics in single trials that correlates with the stages of decision making.

1.3 Overview of thesis

Chapter 2 presents neural data analysis and an accompanying spike-to-fluorescence model that elucidates the difference of neural dynamics when interpreting data from different simultaneous recording technologies, i.e. multi-electrode array and optical imaging. We analyze a large set of recordings performed in the same delayed-discrimination task, under the same conditions, in the same lab with multiple recording approaches, including electrophysiological recordings and imaging. We directly compare the results of a substantial set of typically used analyses and test for any differences. We find several discrepancies at both the single neuron and at population level analyses. Utilizing an additional set of simultaneous imaging and single cell recordings, we construct a phenomenological forward spike-to-fluorescence model that transforms electrophysiological recordings into synthetic imaging data. We use this model to show that many of the differences between the analyses can be attributed to the indirect reporting of neural activity. In our study, spike inference algorithms were

only able to partially undo these differences. These findings clarify the manner by which results from electrophysiological and imaging studies of the same brain area can be compared, and highlight the importance of further understanding the transformations associated with indirect recordings of activity by collecting more ground-truth data and developing further statistical approaches.

Chapter 3 presents a latent variable based analysis of simultaneously recorded neural data to uncover the neural dynamics in the decision-making task. We extend the linear dynamical systems to a time-varying version, which can explicitly incorporate our knowledge of the behavioral epoch (such as sensory sampling and choice memory). We find that the local neural circuit (e.g. that of tens of neurons) for motor planning can provide a continuous neural signal underlying the internal states in different stages of a decision. We identify such a signal in a latent variable space, where each neural mode presents a source of input shared by multiple neurons. The neural-mode dynamics in this shared-input space can predict multiple asprcts of behavioral variability, such as trial type, reaction time, and trial correctness, in single trials, and be robustly maintained for seconds post-decision, and thus represent a single-trial correlate of these properties. Moreover, we confirm that the computational principles of choice formation (and memory) follows the non-leaky drift diffusion model at the population level.

Chapter 4 presents a spiking neuron based model underlying the computation of choice confidence in a local neural circuit of decision making (in comparison with

the experimental observations) (Wei and Wang, 2015). The model is endowed with a continuous network of neurons that can represent any direction; therefore it can be readily extended to incorporate the presentation of a third "sure" target during a delay period. Notably, such a model of decision-making and memory processes was not originally designed for the experiment modeled in this paper (the Kiani-Shadlen experiment (Kiani and Shadlen, 2009)) which attempts to account for confidence estimation. It is thus surprising that the model can capture both a range of behavioral performance data and physiological observations from single neurons in the lateral intraparietal cortex. We noted that neurons selected for the sure target win the competition when the activities of neurons selected for the two alternative choices are indistinguishable. Quantitatively, we found that confidence could be estimated, at any time, as a sigmoid function of the differential firing activity of the two competing neural pools selected for the alternative choices. Therefore, choice confidence is computed simultaneously when a decision is made, and a trial-by-trial variation of choice is generated by sampling of stochastic neural dynamics.

Chapter 5 discusses overall conclusions of this work and highlights future directions to expand our studies. We list some open questions and foreseeable obstacles towards a complete understanding of population codes in the brain. We suggest general approaches for solving these questions.

Chapters 2 and **3** are accompanied by brief technical references for the statistical learning analyses used in both chapters. **Appendix A** first provides a simplified

introduction to sparse linear discriminant analysis (Guo et al., 2007), the main decodability analysis of trial types used in our population analysis. Secondly, **Appendix A** also describes in detail the latent variable model, the time-varying linear dynamical system, which we developed for the analysis of the population code in behavioral state relevant conditions, used in **Chapter 3**.

Chapter 2

A direct comparison of neural dynamics measured with extracellular recording and calcium imaging

Calcium imaging using fluorescent protein sensors is a powerful method to record activity in neuronal populations. However, the relationship between calcium-related fluorescence and spike rates is non-linear and unknown for any one neuron. Here, we compare spike trains and neuronal fluorescence, recorded from matched populations of motor cortex neurons in behaving mice. The slow and variable kinetics of fluorescence obscure rapid changes in neuronal selectivity and dispersed dynamics in time, so that

activity in populations of neurons appeared to tile time. Since calcium imaging is more sensitive to increases rather than decreases in spike rate, measures of neural selectivity were distorted. We used a model of spike-to-fluorescence coupling to transform spike trains into synthetic imaging data. The synthetic imaging data recapitulated the biases seen in actual imaging data. These confounds need to be considered when interpreting measurements of neural activity based on calcium imaging.

2.1 Introduction

Extracellular single unit recordings (hereafter abbreviated as 'ephys') and calcium imaging offer different tradeoffs for interrogating neural populations (**Figures 2.1B-C**). Ephys directly reports the spiking activity of neurons with a high signal-to-noise ratio, temporal fidelity, and dynamic range, but typically offers access only to a sparse subset of relatively active neurons in a local circuit (**Figure 2.1E**) (Buzsaki, 2004). In addition, the ability to track the same population of neurons across time, important for understanding the neural basis of learning, remains challenging (Tolias et al., 2007; Ganguly and Carmena, 2009). In contrast, calcium imaging reports spiking activity indirectly (Grienberger and Konnerth, 2012; Peron et al., 2015). The transformation from spikes to calcium is non-linear because of the dynamics of the intracellular calcium concentration (Scheuss et al., 2006) and nonlinearities imposed by the use of protein-based indicators (Tian et al., 2009; Chen et al., 2013; Akerboom et al., 2012;

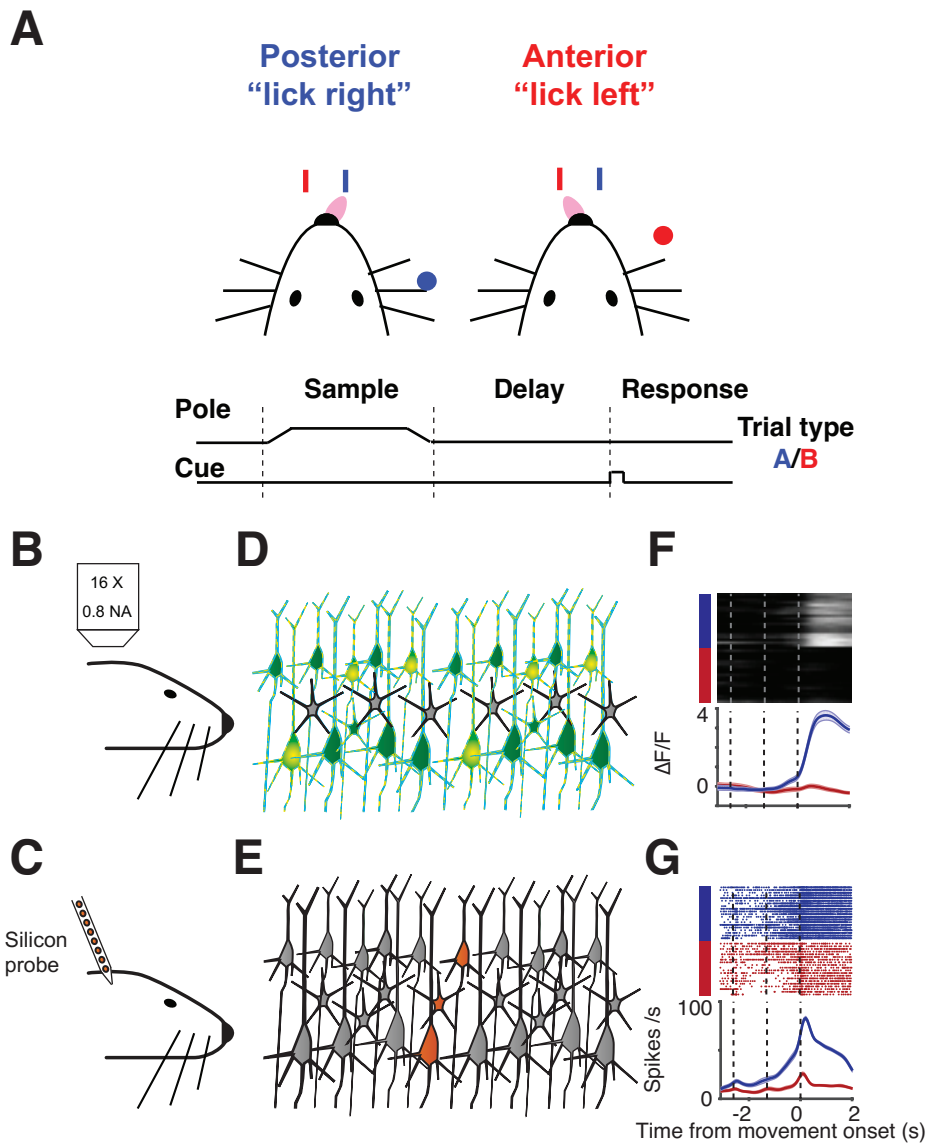
Pologruto et al., 2004). In addition, calcium imaging has limited signal-to-noise ratio for detecting spikes and limited dynamic range (Peron et al., 2015). However, calcium imaging provides access to large numbers of neurons simultaneously (Peron et al., 2015; Huber et al., 2012; Sofroniew et al., 2016), potentially with cell type specificity (Fu et al., 2014; Peron et al., 2015) (**Figure 2.1D**). Moreover, calcium imaging can track the activity of the same neuronal populations over time (Huber et al., 2012; Peters et al., 2014). With the development of highly sensitive fluorescent protein-based indicators (Tian et al., 2009; Chen et al., 2013; Akerboom et al., 2012; Pologruto et al., 2004; Ohkura et al., 2012; Inoue et al., 2015; Dana et al., 2016) and powerful new imaging methods (Sofroniew et al., 2016) calcium imaging has been rapidly adopted for measurements of neural population activity.

During animal behavior, spike rates can vary by orders of magnitude across behavioral epochs and across neurons (O'Connor et al., 2010; Hromádka et al., 2008; Li et al., 2015). Spike rates can change over time-scales from milliseconds to seconds (Brody et al., 2003; Li et al., 2015; Li et al., 2016). In addition, the coupling between individual spikes and calcium-dependent fluorescence changes differs across individual neurons. Many uncertainties remain about how calcium imaging transforms spike trains under these condition. It is unclear what limitations calcium imaging imposes in terms of the analyses that are typically applied to measurements of population activity and the conclusions that can be drawn from the data.

We consider this problem in a challenging context, where the dynamics of the

neural circuit are rich and variable across neurons. Neurons in frontal cortex fire at a wide range of spike rates and exhibit diverse temporal dynamics and selectivity, correlated with behavioral parameters (Brody et al., 2003; Li et al., 2015). Understanding the neural coding underlying the function of these brain regions often takes the form of population analyses (Cunningham and Yu, 2014). These analyses themselves are likely affected by the mode of recording. We analyze ephys and calcium measured in matched neuronal populations in the same behavioral task. We directly compare the results of standard measurements of selectivity and population dynamics. We detect quantitative and qualitative discrepancies at both the level of single cells and neural populations.

Using an additional set of simultaneous imaging and single cell recording data we constrain a phenomenological model that transforms spike trains to synthetic imaging data. We then apply this model to the ephys data to generate synthetic imaging data. Comparing this synthetic data to the recorded imaging data, we find that most of the differences between the analyses can be attributed to indirect and non-linear reporting of neural activity. Spike inference algorithms and other deconvolution methods were of limited use in undoing these differences. Overall our results reveal limitations of calcium imaging as a probe of neural circuit dynamics and highlight the importance of a deeper understanding of the transformation imposed by calcium imaging. More quantitative interpretation of calcium imaging and full utilization of all its advantages will require investment in ground-truth data sets and new statistical approaches.


A
Posterior
"lick right"
Anterior
"lick left"
Sample
Delay
Response
Pole
Cue
Trial type
A/B
B
16 X
0.8 NA
D
F
ΔF/F
4
0
C
Silicon
probe
E
G
Spikes /s
100
0
-2
0
2
Time from movement onset (s)

2.2 Results

Neural activity was measured using electrophysiology (ephys) and calcium imaging under identical conditions. Mice performed a tactile delayed response task (Guo et al., 2014b; Guo et al., 2014a; Li et al., 2015) (**Figure 2.1A**). In each trial, mice judged the location of an object with their whiskers. During the subsequent delay epoch (approximately 1.3 seconds), mice maintained a memory of the previous sensory experience and planned an upcoming response. Following an auditory "go" cue, mice reported object location with directional licking. To emphasize generality, we refer to the posterior pole position trials and their associated lick right instruction as "trial type A" and the anterior pole positions and their associated instruction to lick left as "trial type B".

Calcium imaging and electrophysiological recordings were performed in the left anterior lateral motor cortices (ALM) in separate mice. Imaging was performed us-

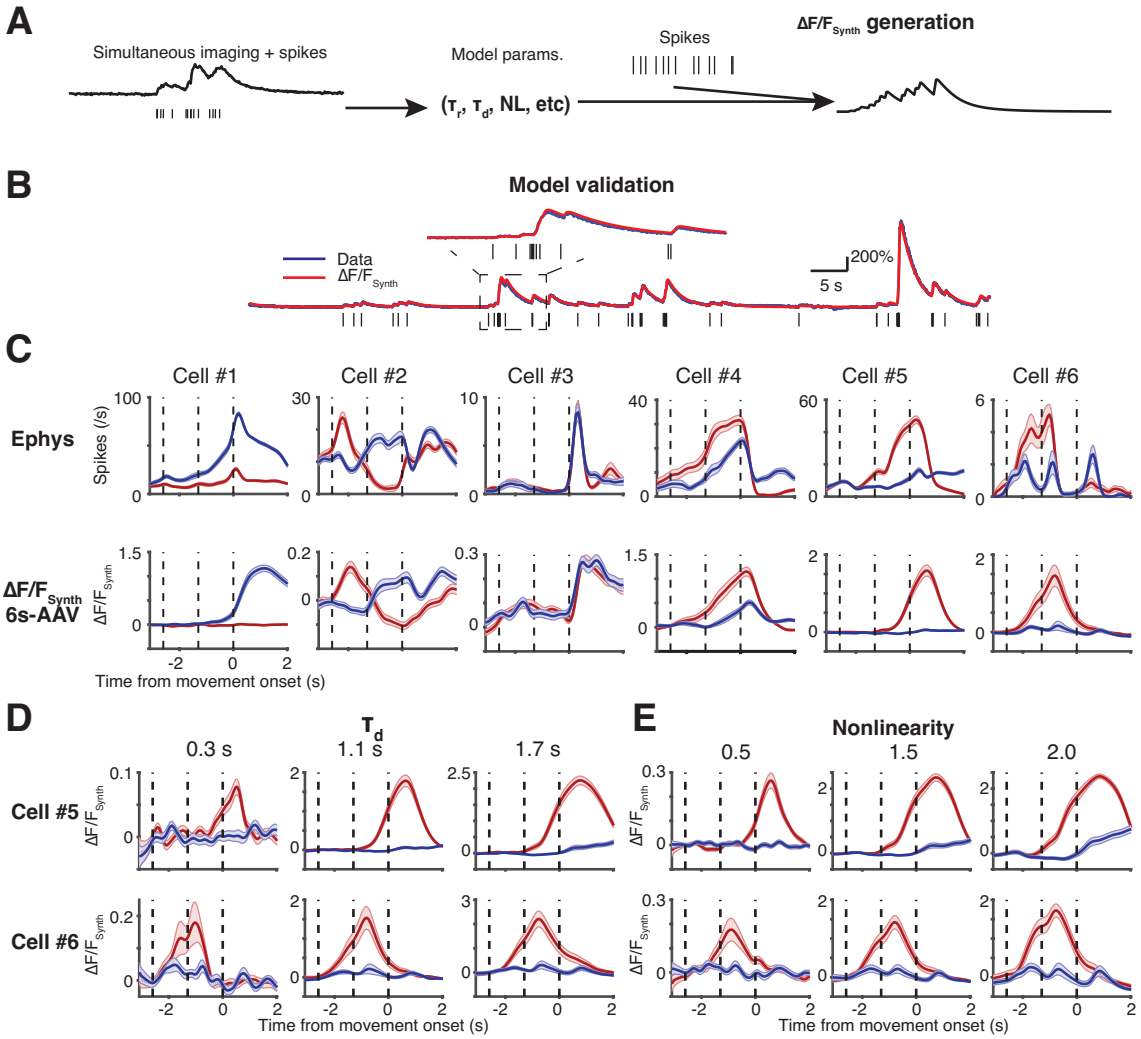
---

Figure 2.1 *(preceding page)*: Simultaneous recordings of ALM population using different techniques.
(**A**) Mice were trained on a delayed-response two alternative forced choice task. Mice discriminated a pole position (anterior or posterior) and reported it by directional licking (lick right, trial A, blue; lick left, trial B, red) in a response to an auditory cue after a delay period. (**B**) Schematic of imaging experiment. (**C**) Schematic description of electrophysiological experiment. (**D**) Schematic of imaging being able to report the activity of hundreds of units (cells in green), and with potential cell-type specificity in recording. (**E**) Schematic of ephys recordings reporting activity from a handful of units at a time (orange recorded units). (**F**) Example dynamics of neuron recorded by imaging with the GCaMP6s calcium indicator (mean activity, thick line; sem, shaded area). (**G**) Example dynamics from a neuron recorded by ephys (mean activity, thick line; sem, shaded area; vertical dash lines indicate switching times of behavioral epochs, time zero is the movement onset time).

ing standard two-photon microscopy (Peron and Svoboda, 2010) (Figure 1B). In one series of imaging experiments neurons were transduced with adeno-associated virus expressing GCaMP6s (6s-AAV), a widely used method that produces robust GCaMP6s expression levels (Huber et al., 2012) (data from (Li et al., 2015), 1493 neurons, 4 mice) (**Figures 2.1B, D, F**). In other experiments we imaged ALM neurons in transgenic mice expressing GCaMP6s in a large fraction of cortical pyramidal neurons (GP4.3, 2293 neurons, 1 mouse). These neurons have lower GCaMP6s expression levels and faster fluorescence dynamics compared to neurons transduced with AAV (Dana et al., 2014). Extracellular spikes were recorded with silicon probes in ALM (720 neurons recorded in 19 mice; **Figures 2.1C, E, G**). The mean spike rate was 5.23 ± 5.76 Hz (mean±std., range 0.26—53.74 Hz). The recordings from (Li et al., 2015) were subsampled so that the distribution of recording depths was similar for imaging (120—740 μm) and ephys (100—800 μm).

2.2.1 A 'spike to calcium' (S2C) model

To compare imaging and ephys, we developed a phenomenological model that converts spike times to synthetic fluorescence time series (Akerboom et al., 2012; Chen et al., 2013; Yasuda et al., 2004; Li et al., 2015). The relationship between spikes and changes in fluorescence is complex, including multiple levels of non-linear effects. Under physiological conditions the change in cytoplasmic calcium concentration per action potential $\Delta[Ca^{++}]_{AP}$ and the calcium extrusion rate both change with $[Ca^{++}]$


A
Simultaneous imaging + spikes
Model params.
(τr, τd, NL, etc)
Spikes
ΔF/FSynth generation
B
Model validation
Data
ΔF/FSynth
200%
5 s
C
Cell #1
Cell #2
Cell #3
Cell #4
Cell #5
Cell #6
Ephys
Spikes (/s)
ΔF/FSynth 6s-AAV
ΔF/FSynth
Time from movement onset (s)
D
Td
0.3 s
1.1 s
1.7 s
Cell #5
Cell #6
ΔF/FSynth
Time from movement onset (s)
E
Nonlinearity
0.5
1.5
2.0
ΔF/FSynth
Time from movement onset (s)

(Scheuss et al., 2006). Down-stream of calcium, protein reporters of calcium (e.g. GCaMP) respond relatively slowly (response time on the order of 100 ms) and non-linearly to $\Delta[Ca^{++}]_{AP}$. Rather than attempting a detailed biophysical model, we fit calibration data with a minimal 5-parameter phenomenological model (S2C). The parameters include a rise-time (τ_r), a decay-time (τ_d), a non-linearity parameter (NL), a half-activation parameter (EC50) and a maximum possible fluorescence change (F_m) (**Materials and Methods**).

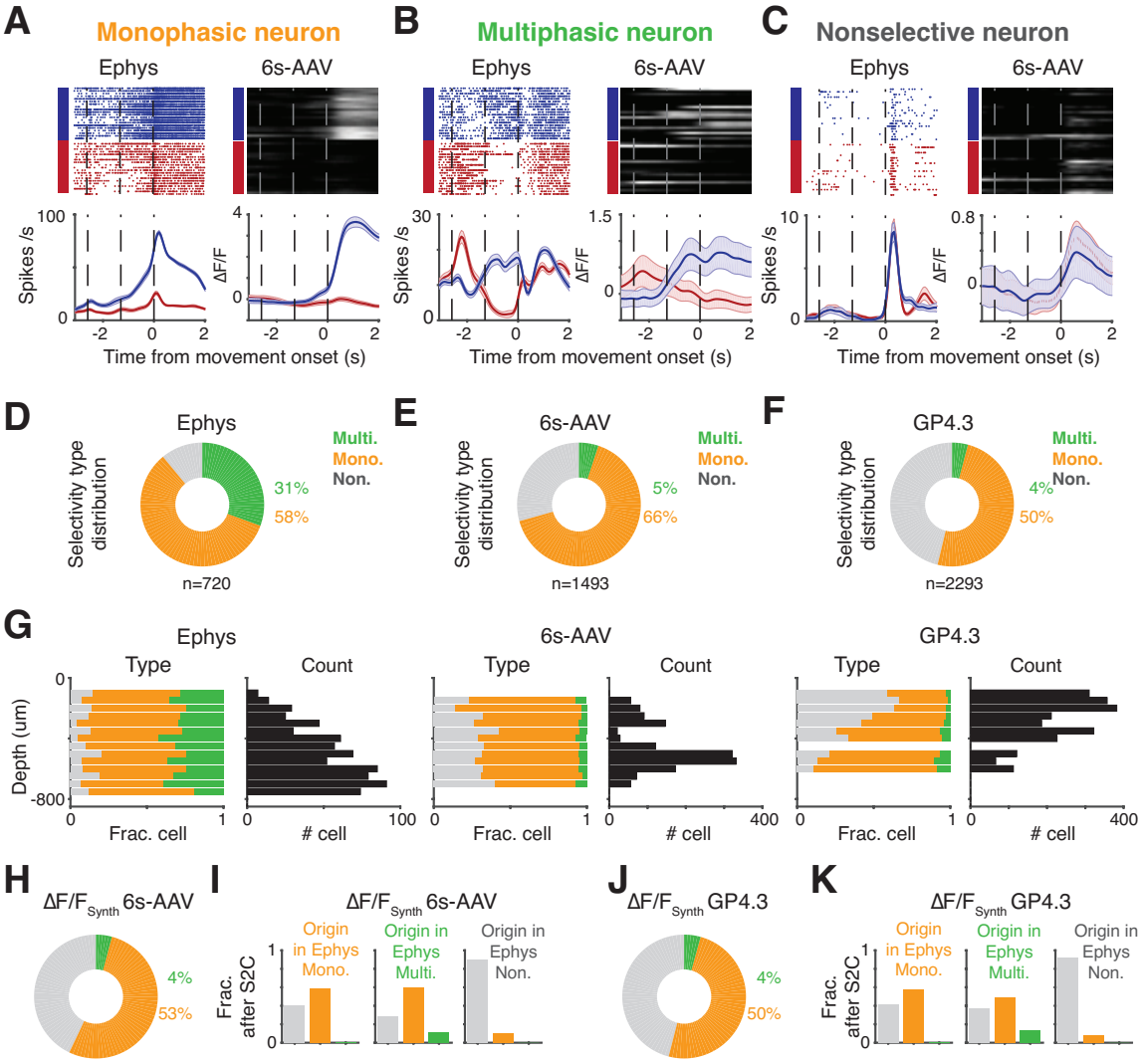
We estimated S2C model parameters from simultaneous loose-seal electrophysiological recordings and imaging (6s-AAV: 9 cells, 21 recording sessions (Chen et al., 2013); GP4.3: 22 cells, 33 recording sessions; **Figures 2.2A, 2.S1A**). The model

---

Figure 2.2 *(preceding page)*: Schematic description of spike-to-calcium model.
(**A**) A phenomenal spike-to-calcium (S2C) model generates a synthetic fluorescence trace from spike events, using a 5-parameter phenomenological model (S2C; Equations 1, 2; **Materials and Methods**). The parameters include a rise-time (τ_r), a decay-time (τ_d), a non-linearity parameter (NL), a half-activation parameter (EC50) and a maximum possible fluorescence change (F_m). The model was first fitted using simultaneous recordings of ephys and imaging, and then applied to spike trains such as to generate the estimation of synthetic fluorescence dynamics ($\Delta F/F_{Synth}$). (**B**) Example fit of fluorescence dynamics from spikes in a GCaMP6s expressing neuron using S2C model. Measured $\Delta F/F$ (blue), simultaneously recorded spikes (black) and simulated $\Delta F/F_{Synth}$ (red) from S2C model (top, a zoomed-in version of local time points). (**C**) Dynamics of 6 example cells in ephys and their S2C model generated synthetic 6s-AAV imaging dynamics. Ephys, top row; corresponding synthetic imaging data, bottom row; each column corresponds to a single neuron. (**D**) Synthetic imaging data for different model parameter values for two example neurons. Length of decay constant is varied across columns; top, Cell #5; bottom, Cell #6. (**E**) Synthetic imaging data for different model parameter values for two example neurons. Slope of nonlinearity is varied across columns; top, Cell #5; bottom, Cell #6.

typically fitted $> 60\%$ of the explained variance in the raw fluorescence dynamics (Given the criterion, 19/21 6s-AAV neurons and 28/33 GP4.3 neurons passed, **Figure 2.S1B**).

We compared spike rates and synthetic fluorescence dynamics ($\Delta F/F_{Synth}$) based on the S2C model (**Figure 2.2C**). The sign of the selectivity (separation of responses across conditions) was typically conserved in the $\Delta F/F_{Synth}$. However, the S2C transformation distorted a variety of neuronal response properties. For example, the dynamics of selectivity was much slower in $\Delta F/F_{Synth}$ compared to ephys (for example Cell#1), which also obscures changes in selectivity (late response epoch in Cell #5, #6). In cases where changes in selectivity were detected, they occurred much later in $\Delta F/F_{Synth}$ compared to ephys, often in adjacent epochs. These biases are caused in part by temporal integration of activity in $\Delta F/F_{Synth}$, which is expected to hide rapid changes in the signal, and by nonlinearities. Furthermore, similar dynamics in $\Delta F/F_{Synth}$ were seen to arise from substantially different spike rate changes (Cell #5, #6). Conversely, the same spike rate change can result in substantially different $\Delta F/F_{Synth}$, given different model parameters within the range of experimentally observed parameters (**Figures 2.2D-E**). Similar biases were seen over the entire range of parameters.


A
Monophasic neuron
Ephys
6s-AAV
Spikes /s
ΔF/F
Time from movement onset (s)
B
Multiphasic neuron
Ephys
6s-AAV
Spikes /s
ΔF/F
Time from movement onset (s)
C
Nonselective neuron
Ephys
6s-AAV
Spikes /s
ΔF/F
Time from movement onset (s)
D
Ephys
Selectivity type distribution
Multi.
Mono.
Non.
31%
58%
n=720
E
6s-AAV
Selectivity type distribution
Multi.
Mono.
Non.
5%
66%
n=1493
F
GP4.3
Selectivity type distribution
Multi.
Mono.
Non.
4%
50%
n=2293
G
Ephys
Type
Count
Depth (um)
Frac. cell
cell
6s-AAV
Type
Count
Frac. cell
cell
GP4.3
Type
Count
Frac. cell
cell
H
ΔF/F Synth 6s-AAV
4%
53%
I
ΔF/F Synth 6s-AAV
Frac. after S2C
Origin in Ephys Mono.
Origin in Ephys Multi.
Origin in Ephys Non.
J
ΔF/F Synth GP4.3
4%
50%
K
ΔF/F Synth GP4.3
Frac. after S2C
Origin in Ephys Mono.
Origin in Ephys Multi.
Origin in Ephys Non.

2.2.2 Hidden dynamics in calcium imaging

Individual ALM neurons exhibit diverse temporal dynamics and even change selectivity across trial epochs (**Figure 2.S2A**) (Guo et al., 2014b; Li et al., 2015). We classified dynamics into three broad categories: "Monophasic" neurons showed consistent selectivity across the trial (**Figure 2.3A**); "multiphasic" neurons changed selectivity over time (for more than 335 ms) (**Figure 2.3B**); "non-selective" neurons responded similarly across trial types but in a task-modulated manner (**Figure 2.3C**). In the electrophysiology data set the majority of neurons were selective (641/720; corresponding to 89%), with a substantial proportion of multiphasic neurons (220/720;

Figure 2.3 *(preceding page)*: Neuronal selectivity measured by calcium imaging exhibits less heterogeneity in temporal dynamics than when measured by electrophysiology.
(**A-C**) ALM neurons exhibit diverse dynamics and trial type selectivity. (**A**) Example monophasic selective neurons; neurons show the same polarity of selectivity in time, from ephys (left; top: raster plots, bottom: mean activity across trial type) and imaging (right; 6s-AAV; top: trial by trial activity, bottom: trial averaged) recordings. (**B**) Example multiphasic selective neurons; neurons show a sustained switch of selectivity over time. (**C**) Example non-selective neurons; neurons show similar dynamics across behavioral conditions. (**D-F**) Fraction of selective neurons in ephys (**D**), 6s-AAV (**E**) and GP4.3 (**F**) imaging. Fractions of mono- (orange), multiphasic (green) and nonselective (gray) cells are shown as donut plot. (**G**) Fraction of multi-, monophasic and non-selective neurons as a function of recording depth for ephys (left), 6s-AAV (middle) and GP 4.3 (right). Each plot shows fraction on left and number of cells on right. (**H**) Same plot as **D** but for 6s-AAV synthetic imaging data. (**I**) Breakdown of how each response type in ephys was transformed to 6s-AAV synthetic imaging after the S2C model. Bars show fraction of neurons that transformed into each type (gray, non-selective; orange, monophasic; green, multiphasic). Left plot shows cells that were monophasic in ephys, middle shows cells that were multiphasic, and right shows cells that were non-selective. (**J**) Same plot as D but for GP4.3 synthetic imaging data. (**K**) Same plot as I but for GP4.3 synthetic imaging data.

corresponding to 31%) (**Figure 2.3D**).

In the imaging data set the proportions of neurons falling into these categories was substantially different. Fewer neurons showed multiphasic selectivity (6s-AAV, 76/1493, corresponding to 5%; GP4.3, 98/2293, corresponding to 4%; $p < .001$, t test, different for both imaging conditions compared to ephys). This occurred despite a similar proportion of monophasic selective neurons (58% Ephys, 66% 6s-AAV, 50% GP4.3). Similar effects were seen across different imaging depths (**Figure 2.3G**).

We used the S2C model to gain an intuition for these biases. For each neuron we computed $\Delta F/F_{Synth}$ with model parameters sampled randomly from the distribution (**Materials and Methods**). We then performed the same population analysis. The S2C transformation caused a significant loss in the number of multiphasic neurons, similar to the imaging data sets (proportion of multiphasic neurons: ephys, 31%; 6s-AAV, 5%; GP4.3, 4%; S2C 6s-AAV, 4%; S2C GP4.3, 4%; two-tail t test, $p < .001$; **Figures 2.3H-J**). Multiphasic neurons became monophasic or non-selective neurons after the S2C transformation (**Figures 2.3I-K**). Therefore, neural dynamics measured with calcium imaging often miss rapid changes in selectivity.

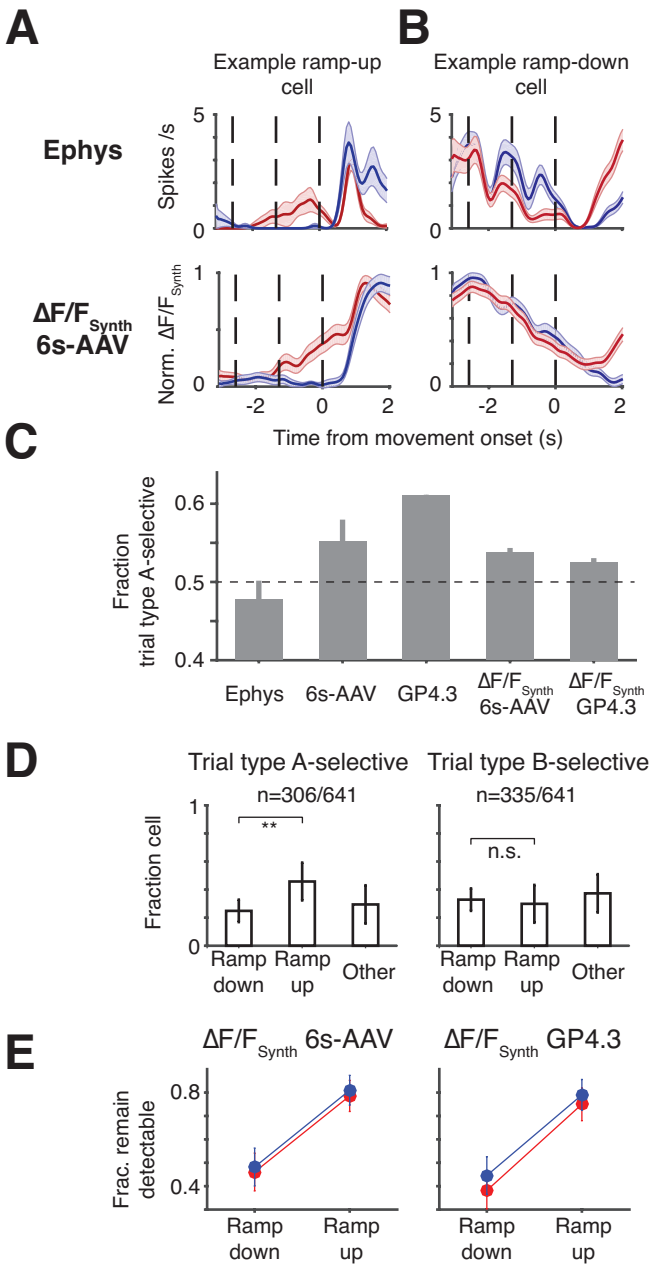
We tested to what extent these biases can be undone by running spike inference algorithms on the calcium data (Calcium-to-Spike, C2S). We used both a direct deconvolution approach and a state-of-the-art C2S model (Pnevmatikakis et al., 2014a; Pnevmatikakis et al., 2016). We find that in both cases spike inference was able to undo only some of the bias. In our hands, the direct deconvolution approach was

more effective and recovered approximately half of the difference in multiphasic selectivity between the ephys and imaging datasets (6s-AAV, 15%; GP4.3, 19%, compare to 31% in ephys; **Figure 2.S3E**). Since we do not have ground truth for this case we can only assume that the fraction should be similar to that in the ephys dataset. Therefore, we additionally ran the C2S models on the $\Delta F/F_{Synth}$ data, where we know the original selectivity of each neuron and can thereby more precisely assess the recovery of dynamics. We find similar results (S2C 6s-AAV, 19%; S2C GP4.3, 18%; **Figure 2.S3H**).

2.2.3 Biased selectivity in calcium imaging

The distortion introduced by calcium imaging depends on the detailed features of the spike rate modulation. If responses to different stimuli have different dynamics, their unequal processing by calcium and calcium reporter may lead to unequal biases. Here we illustrate this point using an analysis of trial type selectivity. ALM spike rates can increase or decrease during the behavioral trial, depending on the trial conditions. We refer to cells with positive modulation as ramp-up cells (**Figure 2.4A**), and cells with negative modulation as ramp-down cells (**Figure 2.4B**) (cells with more complex response patterns were classified as "other" in this analysis; **Figure 2.4D**). We analyzed how this modulation affects selectivity measured with imaging or after the S2C transformation.

Even for neurons with similar spike count differences across trial types, we find a


A
B
Example ramp-up cell
Example ramp-down cell
Ephys
Spikes /s
5
0
ΔF/F Synth 6s-AAV
Norm. ΔF/F Synth
1
0
-2
0
2
Time from movement onset (s)
C
Fraction trial type A-selective
0.6
0.5
0.4
Ephys
6s-AAV
GP4.3
ΔF/F Synth 6s-AAV
ΔF/F Synth GP4.3
D
Trial type A-selective
Trial type B-selective
n=306/641
n=335/641
Fraction cell
**
n.s.
Ramp down
Ramp up
Other
E
ΔF/F Synth 6s-AAV
ΔF/F Synth GP4.3
Frac. remain detectable
0.8
0.4
Ramp down
Ramp up

differential processing of selectivity in $\Delta F/F_{Synth}$ (**Figures 2.4A-B**). For ramp-up cells the separation of activity across trial types was retained in $\Delta F/F_{Synth}$, albeit with slower dynamics (**Figure 2.4A**). In contrast, for ramp-down cells, $\Delta F/F_{Synth}$ did not show selectivity. In other words, selectivity was often conserved in $\Delta F/F_{Synth}$ for ramp-up cells but not for ramp-down cells (**Figure 2.4E**).

This effect introduced biases in calcium imaging data at the level of neural populations. Neurons with type A (lick-right) selectivity and type B (lick-left) selectivity were balanced in the electrophysiology data (Guo et al., 2014b; Li et al., 2015), but showed a type A bias in the imaging data ($p < 0.001$) (**Figure 2.4C**). $\Delta F/F_{Synth}$ showed a similar type A bias ($p < .001$). In the ephys data set, neurons with type A selectivity were more likely to be ramp-up cells, whereas neurons with type B selectivity were more likely ramp-down cells. Since calcium imaging is more likely to capture ramp-up selectivity than ramp-down selectivity, the imaging data picks up a trial type A bias not presented in the ephys data set.

Figure 2.4 *(preceding page)*: Induction of distinct activity dependent biases in different populations of neurons by calcium dynamics.
(**A-B**) S2C model predicts that selectivity of ramp-down cells, comparing to that of ramp-up cells, would be hard to detect in imaging. (**A**) An example ramp-up cell (top, neural dynamics in ephys; bottom, dynamics in S2C 6s-AAV), selectivity remains detectable in synthetic imaging data. (**B**) An example ramp-down cell (top, neural dynamics in ephys; bottom, dynamics in S2C 6s-AAV), selectivity becomes undetectable in synthetic imaging. (**C**) Fraction of contra-selective neurons (those with enhanced response for anterior pole position) in the different datasets. (**D**) Bar plot of fractions of ramp-up, ramp-down and "other" cells in ephys for contra-selective (left) and ipsi-selective (right). (**E**) Fraction of cells that remain selective in synthetic imaging (S2C 6s-AAV, left; S2C GP4.3, right) for ipsi-selective (blue) and contra-selective (red) cells, separately for ramp-up and ramp-down cells.

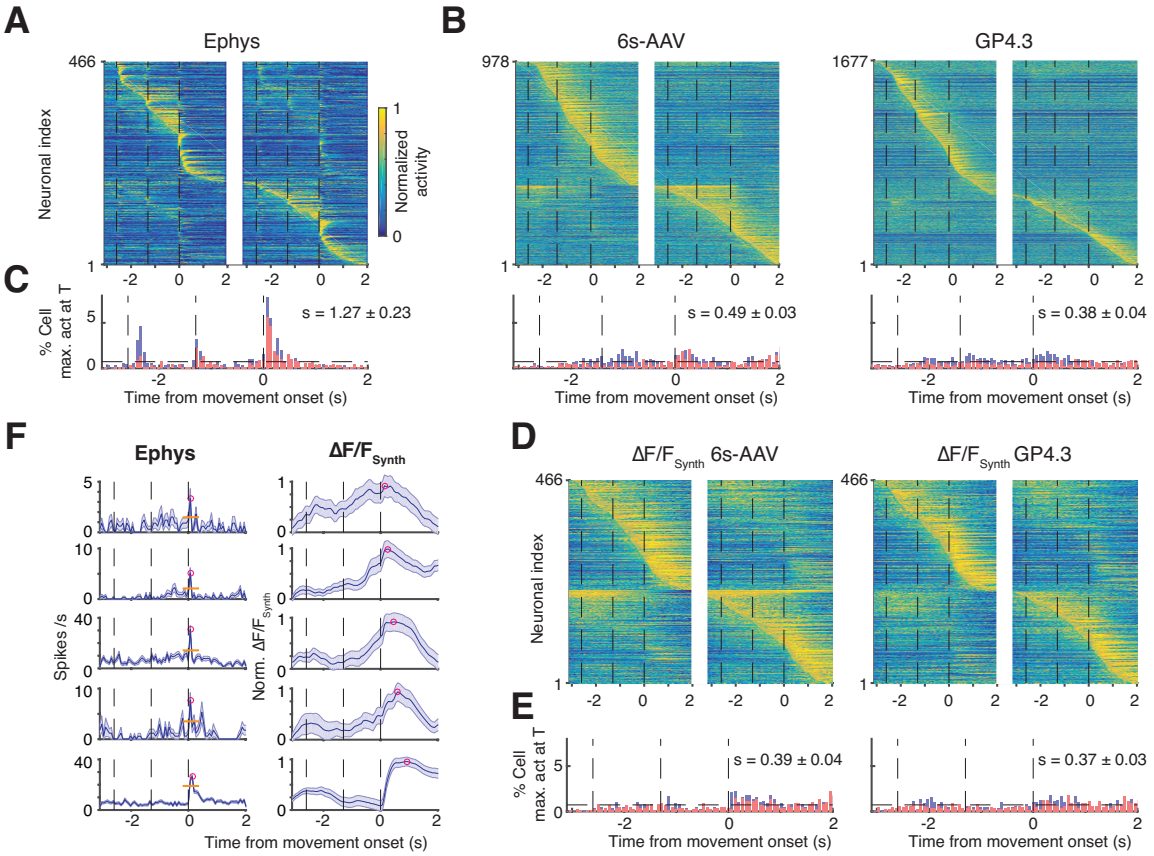


Figure 2.5: Calcium imaging exhibits more sequence-like population activity than that of ephys recordings.
(**A**) Heatmap of normalized trial-averaged firing rates for lick left trials (left) and lick right trials (right) for ephys data. Firing rates were normalized to maximum of activity across both conditions and neurons were sorted by latency of peak activity and by their preferred trial type. (**B**) Same plots as **A** but for 6s-AAV (left) and GP4.3 (right). (**C**) Fraction of neurons with a peak at given time point over time. distribution in time (ephys, left; 6s-AAV, middle; GP4.3, right) plotted simultaneously for both trial types (red: right preferring trials, blue: left preferring trials, black horizontal line: uniform distribution). (**D-E**) The same plots as **B-C** for synthetic imaging (S2C 6s-AAV, left; S2C GP4.3, right). (**F**) Example cells with peaks at a similar time in ephys (left; mean activity, thick blue line; sem, shaded area; peak, magenta circle; baseline, orange thin line) along with the corresponding synthetic data (right). Neurons are sorted according to their peak times in synthetic imaging (early to late, from top to bottom).

2.2.4 Distortion of network dynamics in calcium imaging

Neurons show temporally complex responses, even in simple trial-based behaviors (Romo et al., 1999; Brody et al., 2003; Rigotti et al., 2013). The details of these spike rate changes are critical for an understanding of circuit models of neural computation. The slow and nonlinear dynamics of calcium imaging could lead to a distorted view of neuronal dynamics. The spike rates recorded in ALM preferentially exhibit peaks in activity at the transitions between behavioral epochs (**Figure 2.5A**) (Li et al., 2015; Akhlaghpour et al., 2016). In contrast, in the calcium imaging data, peaks of fluorescence were spread almost uniformly across trial time, producing a sequence-like appearance (**Figure 2.5B**).

We measured the peakiness of the distribution of neuronal activity across recording modalities. We defined the difference, s, between observed neural activity and uniformly distributed neural activity as the integrated difference between the empirical distribution of peak times and that expected from a uniform distribution ($P = \frac{1}{2T}$):

$$s = \frac{1}{P}\sqrt{\frac{1}{2T}\sum_{i=A,B}\int_0^T dt(P_i(t) - P)^2}.$$

s was much larger for the ephys dataset (1.27 ± 0.23) compared to the 6s-AAV (0.49 ± 0.03; one-tail t-test, $p < .001$) and GP4.3 (0.38 ± 0.04; one-tail t-test, $p < .001$) imaging data. $\Delta F/F_{Synth}$ was similar to the imaging data ($s = 0.39 \pm 0.04$, S2C

6s-AAV; $s = 0.37 \pm 0.03$, S2C GP4.3; **Figure 2.5E**).

The temporally dispersed nature of the imaging data in part stems from the variability of calcium indicator dynamics at the single cell level, which transforms a given time of a change in spike rate into a differently timed peak in calcium activity. Moreover, multiple features of the detailed response can affect this temporal shift. One such feature is the ratio of the peak activity to the baseline before the peak. When activity was weak, the peak of $\Delta F/F_{Synth}$ was relatively close to the peak in ephys (**Figure 2.5F**, top). If the baseline was high, the peak of $\Delta F/F_{Synth}$ was more shallow and delayed (**Figure 2.5F**, bottom; **Figure 2.S4A**). This effect did not explain all the variance and additional features of the response affected the shift.

We find that application of spike inference algorithms only reduced this temporal dispersal by a small fraction (imaging data, $s = 0.32 \pm 0.02$, 6s-AAV; $s = 0.28 \pm 0.01$, GP4.3; **Figure 2.S4D**; synthetic imaging data, $s = 0.45 \pm 0.05$, S2C 6s-AAV; $s = 0.38 \pm 0.04$, S2C GP4.3; **Figure 2.S4G**).

2.2.5 Explained variance of temporal dynamics and trial-type selectivity in leading principal components

Large-scale recording methods can simultaneously record the activity of many neurons. Dimensionality reduction techniques are then typically used to provide a

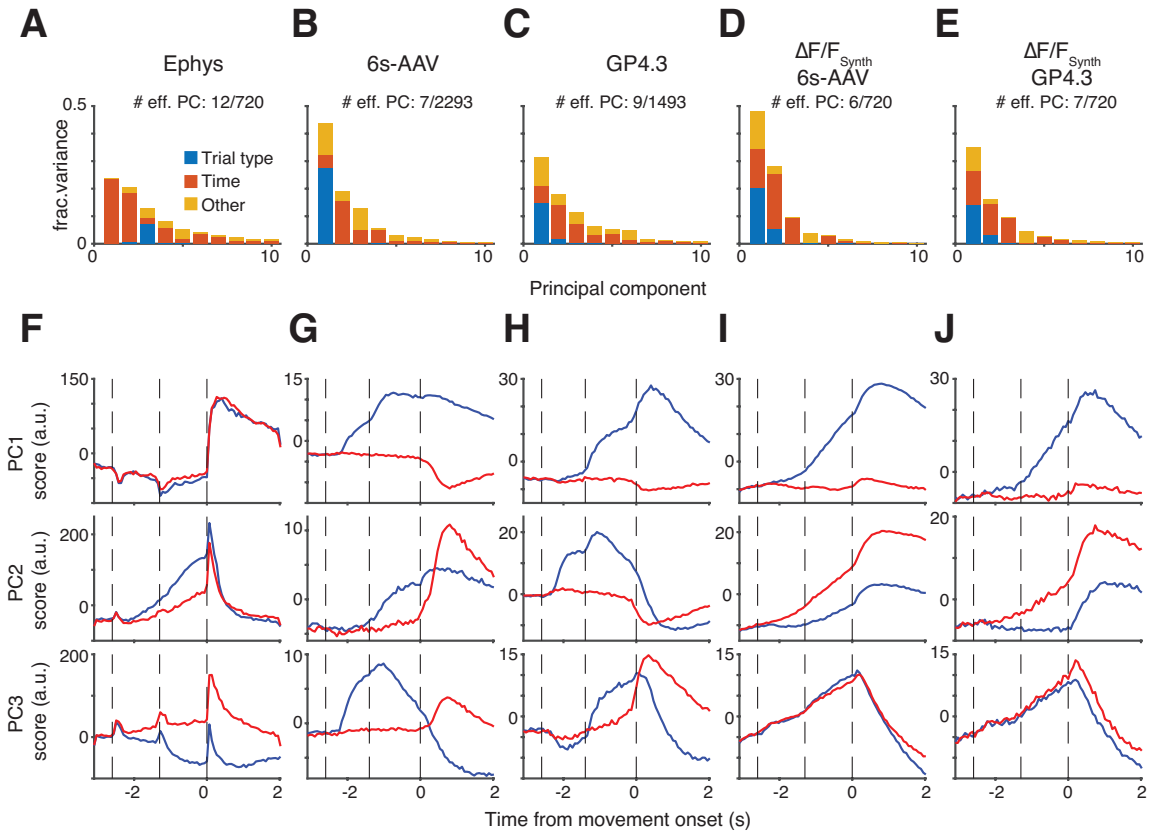


Figure 2.6: Temporal dynamics account for most variance in first principal component of ephys data, but trial type selectivity accounts for most variance in calcium data.

(**A-E**) Percentage of variance of neural activity explained by each principal component (PC; Ephys, **A**; 6s-AAV, **B**; GP4.3, **C**; S2C 6s-AAV, **D**; S2C GP4.3, **E**) shown in bar height. Bar subdivision into colors denotes contents of variability: temporal dynamics (red), trial type (blue) and other factor (yellow, interaction of time and trial type). (**F-J**) Dynamics of first three PCs (from top to bottom) for the two trial types (trial A, blue; trial B, red). Ephys, **F**; 6s-AAV, **G**; GP4.3, **H**; S2C 6s-AAV, **I**; S2C GP4.3, **J**.

compact description of the data (Cunningham and Yu, 2014). For example, PCA is used to find modes of population activity that capture the largest amount of variance in neural activity (Cunningham and Yu, 2014; Kobak et al., 2016) and dynamics are explored in the reduced space. We applied PCA decomposition to the ephys and calcium imaging data. We find that the transformation imposed by calcium imaging produces a qualitatively different PCA decomposition of neural activity.

We rank ordered principal components (PC) that explained more than 1% of variance. In ephys a large number of PCs contributed substantially to the variance, whereas in imaging and synthetic imaging most variance was explained by the first few PCs ($p < .001$; t test, bootstrap). Moreover, the content of the most significant PCs was different between ephys and imaging. Based on the first 10 PCs in ephys and imaging, we estimated the relative contribution to each PC of the temporal dynamics, trial-type selectivity and their interaction (**Materials and Methods**). We found that the fraction of explained variance (EV) due to temporal dynamics was high in the 1st PC in the ephys dataset ($98.71 \pm 0.06\%$, mean±std, 1000 bootstrap), whereas trial-type selectivity was high in the 1st PC of imaging (6s-AAV: 60.39±0.29%; GP4.3: $44.51 \pm 0.65\%$) (**Figures 2.6A-C**, EV; **Figures 2.6F-H**, dynamics). Intuitively, this bias is consistent with the smoothing out of activity reducing the variance contributed by the within-trial dynamics. Similar to imaging, EV of trial-type selectivity was high in the 1st PC of $\Delta F/F_{Synth}$ (S2C 6s-AAV: $44.65 \pm 0.31\%$, **Figures 2.6D, I**; S2C GP4.3: $39.49 \pm 0.61\%$, **Figures 2.6E, J**).

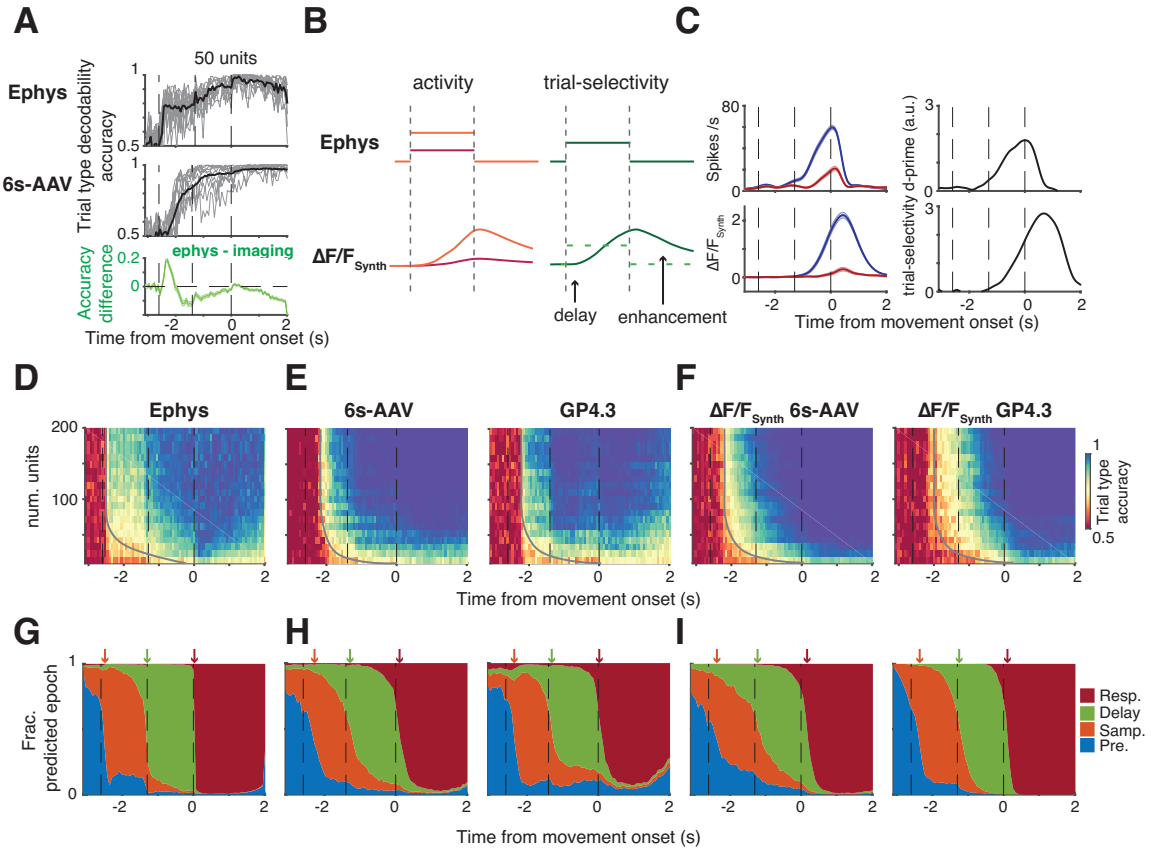


Figure 2.7: Calcium imaging data show a delayed increase of selectivity. (**A**) Performance of instantaneous trial linear-discriminant-analysis (LDA) decoder for 50-unit subpopulations in ephys (top; performance for a random collection of neurons, gray; the average performance across random collections, black thick line) and 6s-AAV (middle) and their difference in time (bottom; green; thick line, mean; shaded area, sem). (**B**) Schematic of relation between activity (left) and decodability (right) in a toy model that has two constant levels of activation for the two trial types (orange and red). (**C**) Example cell showing similar behavior to the toy model. (**D**) Heat map of decodability accuracy as it evolves over time (x-axis) and with increasing neuron number in decoder (y-axis). The gray line in the heatmap presents increasing times of decodability. (**E-F**) Imaging (**E**) and synthetic imaging (**F**) showed a delayed increase in sample epoch and a delayed decay of decodability in response epoch as well as increased accuracy of decoding. (**G**) Performance of time-invariant epoch LDA decoder to behavioral epochs in ephys. The probability of neural activity that LDA decode considers to represent pre-sample (blue), sample (orange), delay (green), and response (red) epoch; arrows indicate the transition times of epochs from neural codes. Ephys exhibited a sharp transition of decodability of epochs, aligned with that of behavior. (**H-I**) The same plot as **G** for imaging (**H**) and synthetic imaging (**I**). Transition times of epochs from neural codes were both delayed in imaging and synthetic imaging.

2.2.6 Calcium indicator dynamics can enhance instantaneous decodability of trial-type variables, but delay the observed response to changes in trial epoch

We found that many cells showed substantial delays in the increase of trial selectivity in the synthetic calcium dynamics (**Figure 2.2C**, Cells #1, #4, and #5). We hypothesized that such a delayed increase of trial selectivity would hold true even at the population level. We measured the instantaneous discriminative power of neuronal activity over time by performing linear discriminant analysis (LDA; **Materials and Methods**). Discrimination was possible even with tens of units and generally increased over time following the beginning of the sample period (Figure 7A). The average decodability in ephys increases earlier (one-tail t-test, $p < .001$), but saturates at a lower level (one-tail t-test, $p < .001$) than that in calcium imaging (**Figure 2.7A**).

A toy model explains both observations (**Figure 2.7B**). Consider a neuron firing at two different levels of activity at a constant rate value during a behavioral epoch (**Figure 2.7B**). For the ephys data its trial-type selectivity is directly proportional to the instantaneous difference of activity (**Figure 2.7B**, top). However, this is not the case in the synthetic data (**Figure 2.7B**, bottom). First, the slowness of

the integration process will result in a delay in the increase of selectivity. Second, intuitively during the period were the neuron is firing selectively, doing instantaneous decoding is non-optimal. Since the signal (difference in activity) is constant and sustained, it would be better to integrate across time and thereby suppress noise. The dynamics of calcium causes that to occur, even if we ostensibly are performing instantaneous decoding on the data. This causes an increase of selectivity in imaging at the peak (**Figure 2.7B**). This effect can be seen at the single cell level (**Figure 2.7C**; delay -2s, ephys; -1.3 s, S2C 6s-AAV; one-tail t-test, $p < .001$; enhancement of selectivity at peak (one-tail t-test, $p < .001$) for this example neuron as well as the population level (**Figures 2.7D-F**).

Performance improved with larger number of units included in analyses (**Figures 2.7D, E**) in both ephys and imaging. Decoding accuracy increased significantly earlier in ephys than imaging (one-tail t-test, 6s-AAV: $p < .001$; GP4.3: $p < .001$) and that later in the response period the instantaneous decoder was more accurate based on imaging than ephys (6s-AAV: $p < .001$; GP4.3: $p < .001$, **Figures 2.7D-E**). Both observations can be explained by the S2C models (delay: one-tail t-test, 6s-AAV, $p < .001$, GP4.3, $p < .001$; enhancement: 6s-AAV, $p < .001$, GP4.3, $p < .001$; **Figure 2.7F**).

Population dynamics vary across time and in particular across behavioral epochs. Which recording method will be more accurately able to track such changes is a priori unclear. The crisp report of dynamics in ephys may favor that method, however if

such changes in dynamics happen in a coordinated manner across the population, the larger number of units typically recorded in imaging may offer a relative advantage. We performed decoding analysis attempting to predict the current behavioral epoch from population activity in order to assay the ability to track network dynamics (**Materials and Methods**). In ephys (**Figure 2.7G**) following a change in behavioral epoch, we observe a rapid decrease of the probability of activity to belong to the previous epoch along with a sharp increase in the probability of belonging to the current epoch. In contrast, in the calcium imaging data such changes tended to be delayed and gradual (**Figure 2.7H**). This effect was recapitulated in the synthetic calcium data from the S2C model (**Figure 2.7I**).

In summary, the analyses of population decodability for trial type and behavioral epoch show a stereotypical shaping of the observed population dynamics by the recording method: the longer integration of the calcium indicator smears changes in dynamics, but perhaps less intuitively constantly integrating a decodable signal yields to higher signal-to-noise ratio in the calcium data. It thus displayed a trade-off between decodability of stimulus and precision of temporal dynamics across recording methods.

2.3 Discussion

Spikes serve as a fundamental currency in communication among neurons and thus a key target for recording neural activity, yet imaging offers many of important opportunities such as larger scale, denser, and specific recordings (Scanziani and Hausser, 2009; Ji et al., 2016; Sofroniew et al., 2016; Vladimirov et al., 2014). Recent advances in the design of calcium indicators with high sensitivity and high reliability show that under conditions of single, or few, action potentials, the firing rate can often be decoded from the calcium imaging (Chen et al., 2013; Dana et al., 2014). Whether this holds in the general case is unclear. In general, there is a paucity of ground truth data, i.e., new simultaneous recordings of spiking and calcium activity from the same neuron. In particular, the data that is available has been recorded only from few brain regions and under limited firing rate conditions (Akerboom et al., 2012; Chen et al., 2013; Li et al., 2015; Greenberg et al., 2008; Grewe et al., 2010; Vogelstein et al., 2010; Vogelstein et al., 2009; Yaksi and Friedrich, 2006; Pnevmatikakis et al., 2014a; Yang et al., 2016; Theis et al., 2016). Accordingly, it is unclear at this point how accurately spike inference can be performed for any particular experiment. It is therefore premature to believe that imaging data can be automatically converted back to ephys and the same analyses typically applied to ephys data can be performed on the inferred data with no additional concerns. Accordingly, we set out to systematically compare different analyses in which ephys and imaging can be applied to interrogate neuronal dynamics.

In our study, we (1) illustrate the main differences in commonly used analyses between two large sets of recording data, one with imaging and one with ephys; (2) relate these biases to our understanding of the dynamics of calcium indicators; and therefore (3) highlight the aspects that techniques to reverse engineer ephys from imaging, or other statistical approaches, may be able to better address in the future.

2.3.1 Biases in imaging dynamics

It is well known that the long effective decay time of calcium dynamics will distort the ability to read out some dynamics in imaging (Li et al., 2015). By first fitting a forward spike-to-calcium model to simultaneously recorded ephys-imaging data, we were able to generate data sets of synthetic imaging data from the non-simultaneous ephys recordings. Comparing these datasets to real imaging data collected under the same conditions, we revealed that many of the discrepancies, e. g., loss of multiphasic selectivity and decrease of peakiness (Harvey et al., 2012; Morcos and Harvey, 2016; Malvache et al., 2016; Picardo et al., 2016) in dynamics, can be accounted for by the synthetic calcium dynamics in the forward model, both in single neuron and population analyses.

Population analyses have become widely used in systems neuroscience as techniques that allow population recordings have become more common (Brown et al., 2004; Cunningham and Yu, 2014; Harris et al., 2016). However, it is unclear whether such analyses would be better served by the crisper recording of dynamics by elec-

trophysiology or by the larger sampling of the network in imaging. More specifically, population analyses often involve dimensionality reduction, e.g., Principal Component Analysis (PCA), to allow for better intuition in interpreting the results (Cunningham and Yu, 2014). Our observation that when using PCA on imaging data, the time-content of the leading Principal Components (PCs) was weakened and the trial-type-content became predominant, is a cautionary note regarding how seemingly identical choices between ephys and imaging, such as performing PCA and keeping the two largest PCs can lead to different aspects of the data retained. In addition, we show that other seemingly identical choices, such as performing the analysis on the same time relative to a transition between behavioral epochs (when the timing of the appearance of a switch in network dynamics is different) or performing instantaneous decoding with the same temporal window (when one recording method effectively integrates the other) can lead to misleading results caused not by neural dynamics, but by the difference in the properties of the recordings.

More generally, any interrogation of the population properties of a complex system is bound to be incomplete and therefore different forms of probing these properties may well over- or under-emphasize different aspects of the data. Choosing what is the best approach will typically depend on the analysis one has in mind. As systems neuroscience involves many types of questions, and there are few standard, agreed upon appropriate analyses, such tradeoffs cannot be determined fully in general. That being said, experimental conditions can ameliorate particular known biases. For

instance, to deal with the delayed report of switches in dynamics, one could use longer behavioral epochs. Indeed, we found more multiphasic neuron in GP4.3 data using 4s delay experiments (**Figure 2.S2B**).

2.3.2 Quantifying the specific variance introduced by indirect measurements of activity

What more could one do with our findings? In addition to just being aware of the biases we point out, one can use our understanding of the relation between ephys and imaging (e.g., forward models) to map out the space of spiking patterns of activity that are consistent with a specific recorded pattern of calcium activity. As we demonstrate in **Figure 2.S3A**, a particular imaging recording could result from a variety of combinations between ephys dynamics and calcium dynamics (i.e. for instance the interaction between the "peakiness" of the particular neurons activity and that neurons particular effective decay time). Not all these possible source combinations will lead to equal differences in the actual value that is extracted by a particular analysis (e.g., the general selectivity of a neuron). Such variability can potentially be quantified by appropriate statistical methods. In other words, a more nuanced view of utilizing inference algorithms to undo biases introduced by indirect recording of activity, is to attempt and directly get a handle on the variance of the summary statistics one is interested in, given the different ways in which an imaging data set

could have arisen from spiking activity.

As more ground truth data is collected using simultaneous-imaging-ephys recordings in brain areas with rich dynamics such as the frontal cortex and as the statistical tools improve, we expect our ability to identify the class of all possible patterns of activity that are consistent with a given imaging dataset and a given noise level to improve. One could then be more quantitative about exactly how much irreducible bias is introduced for a particular analysis on a particular sample of recorded activity.

2.4 Materials and Methods

2.4.1 Electrophysiological and imaging datasets

Mice were trained to perform a delayed version of tactile discrimination task upon the stimulus (pole) position onto the whisker, while electrophysiological (ephys) or calcium imaging recording was performed (Guo et al., 2014b; Guo et al., 2014a; Li et al., 2015). The total duration of sample-delay was set identically to 2.6 s across datasets. In ephys, the delay period was 1.3 s; in both imaging, that was 1.4 s (**Table 2.S1**). Trials with early licking behavior were excluded in analysis. The neurons in ephys were sampled from 100 to 800 um in depth; those in 6s-AAV were within 150|740 um; those in GP4.3 were within 120|640 um. In all datasets, we selected neurons with > 20 trials for each type (A and B). For imaging data, we performed a post-hoc detection of outlier in recordings to remove trials where $> 30\%$ of the

time points contains a signal 3 or more standard deviations away from median. This reduced the total number of neurons with sufficient number of trials, yielding 1493 and 2293 units for 6s-AAV and GP 4.3 imaging, respectively.

2.4.2 Simultaneous electrophysiology-imaging recordings

We used the publicly available datasets provided by the GENIE project, Svoboda lab, at Janelia on http://crcns.org. The data were collected using simultaneous loose-seal electrophysiological recordings and imaging of GCaMP6s-expression neurons in primary visual cortex of mice in two conditions, i.e. 6s-AAV and GP4.3. The electrophysiology recordings were performed identically, while imaging had small differences in two conditions (6s-AAV: 60 Hz imaging, 256 $\times$ 256 pixels, 30 $\times$ 30 μm^2; GP4.3 high-zoom: 55 Hz imaging, 256 $\times$ 256 pixels, 37.5 $\times$ 37.5 μm^2). The detail of 6s-AAV dataset was described in (Chen et al., 2013), that of GP4.3 was in (doi:10.6080/K0S46PV7) (**Table 2.S2**).

2.4.3 Description of the 'spike to calcium' model

We developed a phenomenological model that converts spike times to synthetic fluorescence time series (Chen et al., 2013; Akerboom et al., 2012; Yasuda et al., 2004;

Li et al., 2015). This 'spike-to-calcium' (S2C) model follows two equations:

$$c(t) = \sum_{t>t_k} \exp(-\frac{t-t_k}{\tau_d})[1-\exp(-\frac{t-t_k}{\tau_r})] + n_i(t) \tag{2.1}$$

$$\Delta F/F_{Synth}(t) = \frac{F_m}{1+\exp[-k(c(t)-c_{1/2})]} + n_e(t) \tag{2.2}$$

In **Equation 2.1**, spikes are converted to a latent calcium variable $c(t)$, by convolving the spike times, t_k, with a double-exponential kernel, modeled using a rise time, τ_r, and a decay time, τ_d. Internal Gaussian noise, n_i $N(0, \sigma_i^2)$ (amplitude at σ_i), was added thereafter and negative values were set to zero. We then converted $c(t)$ to a synthetic fluorescence signal through a sigmoidal function (**Equation 2.2**; where a non-linearity parameter, k; a half-activation parameter, $c_{1/2}$; and a maximum possible fluorescence change, F_m) upon which was added external Gaussian noise, n_e $N(0, \sigma_e^2)$ (amplitude at σ_e) (Tsien, 1989; Maravall et al., 2000; Yasuda et al., 2004). We first estimated the parameters in each imaging condition using a collection of simultaneous ephys-imaging recorded GCaMP6s-expressing neurons in two imaging conditions (6s-AAV and GP4.3; **Figure 2.S1A**) and then applied them by randomly sampling from their distributions to generate synthetic calcium imaging counterparts of ephys recordings (S2C 6s-AAV and S2C GP4.3). Due to the mismatch of spike rate in ephys and that in simultaneous ephys-imaging recordings, we refitted the nonlinearity based on non-simultaneous data, by estimating, for each neuron's nonlinearity parameters (while linear parameters were fixed) that would

have led to the closet fit with the best possible partner ($r_s > .7$ for mean activities; in which condition, all cells in ephys could find > 1 partner in imaging) across all imaged cells (each target neuron was used < 5 times to avoid biases) in each imaging condition.

2.4.4 Single neuron analyses

To measure the dynamics of selectivity, we performed two-sample t test with neural activity over finer time scales, 67 ms discrete bin (a single frame duration in imaging) (**Figure 2.S2A**). We define a neuron as monophasic if it had consistent polarity of selectivity ($p < .05$) for > 335 ms (5 continuous frames), a neuron as multiphasic if it had a switch of selectivity ($p < .05$) and periods of neuron being selective were all > 335 ms. The rest of the neurons were considered as nonselective neurons, which showed similar activities across trial types. Selectivity index, d-prime (**Figure 2.7C**) of single neuron was computed as

$$d'(t) = \frac{|R_A(t) - R_B(t)|}{\sqrt{\sigma_A^2(t) + \sigma_B^2(t)}},$$

where $R(t)$ is mean activity at time t for each trial type; $\sigma^2(t)$ is the variance of activity across trials at time t; subscript A and B standards for the trial type.

For selective neurons (mono- and multiphasic neurons), we classified them into trial-A- and trial-B-preferring cells, which had a higher averaged activity in the cor-

responding trial type (**Figure 2.4**). We classified neural dynamics as ramp-down (ramp-up) in ephys as averaged activity in sample-delay epochs is less (greater) than that in pre-sample epoch (paired t-test, $p < .05$ across trials). The rest of unclassified cells (other than ramp-up or down) were named as other, which would ramp up for one trial type and ramp down for the other (**Figure 2.4D**).

2.4.5 Principal component analysis

Principal component (PC) analysis was performed on the activity of neurons averaged across trial type ($s \in \{A, B\}$),

$$\mathbf{r}(s,t) = \mathbf{C}\mathbf{x}(s,t) + < \mathbf{r} >_{s,t},$$

an $n \times 2T$ matrix, where n is the number of recorded units in each datasets; T is the number of time points for each trial type; $< \mathbf{r} >_{s,t}$ is the mean activity across time and trial type; $\mathbf{x}(s,t)$ is an $n \times 2T$ PC score matrix and its ith row stands for the ith PC score. Explained variance (EV) of temporal dynamics $EV_i(t)$ and trial-type selectivity $EV_i(s)$ for the ith principal component (PC) were computed as:

$$EV_i(t) = \frac{<< x_i(s,t) >_s^2 >_t}{< x_i(s,t)^2 >_{t,s}},$$

$$EV_i(s) = \frac{<< x_i(s,t) >_t^2 >_s}{< x_i(s,t)^2 >_{t,s}},$$

respectively.

2.4.6 Population decodability analysis of trial type and behavioral epoch

We applied linear discriminant analysis (LDA) on neural dynamics grouped into single frame bins (67 ms non-overlapped bins) to compute the instantaneous decodability of trial type. The optimal LDA decoder was computed separately for each time bin, using trials in correct responses. We estimated performance of the instantaneous LDA decoder by averaging over 100 samplings of subpopulations of recording units (number of the units ranged from 10 to 200 in each sample; **Figures 2.7D-F**). In each sample, we separated the trials of each neuron into non-overlapped training (70%) and testing (30%) sets. We then shuffled trial identity of each neuron to build up random collections of population activity at each time point for correct trial-A (50% in training and testing) and trial-B using all neurons being recorded. The instantaneous decoder of trial type was computed from training set and its performance was tested by testing set (**Figures 2.7A, D-F**).

We tested the ability of neuronal population activity at different times to discriminate the behavioral epoch by using a four-class LDA (i.e. pre-sample, sample, delay and response epoch). In this analysis, we assumed a single optimal LDA decoder that could predict probability of each behavioral epoch from the instantaneous neuronal

activity. We performed the identical estimation procedure of the optimal LDA decoder to that for trial type (training set) and its prediction of behavioral epoch at each time for each trial (testing set, non-overlapped with training). The probability of each behavioral epoch was then estimated as average across trials from the instantaneous neuronal activity at each time (**Figures 2.7G-I**). To compute the latency of neuronal response to behavioral epoch, we use a threshold, 0.7, to decide whether estimated probability of previous behavioral epoch is small enough to accept a change of epochs (arrows on **Figures 2.7G-I**).

To achieve a robust estimation of coefficients, we applied a sparse version of LDA where the number of non-zero coefficients was minimized (Guo et al., 2007).

2.5 Supplementary

2.5.1 Description of calcium-to-spike models

We validated our conclusion of calcium-to-spike inference using two classes of models. The first class of model is direct deconvolution (a supervised learning like algorithm), where we performed the inverse function from calcium dynamics to ephys (Theis et al., 2016), and the second class is the state-of-art generative model based on Bayesian inference technique (an unsupervised learning based algorithm) developed by Paninski's lab (Pnevmatikakis et al., 2014a; Pnevmatikakis et al., 2016).

1. C2S Precise nonlinear decoder

For synthetic imaging data, $\Delta F/F_{Synth}$, from spike-to-calcium model, we could perform a precise direct inverse function (**Equations 2.3**, and **2.4**) to decode the spike rates in each "frame" (non-overlapped time bin, $\Delta t = 67$ ms), on either single trial or on the average across trials (**Figures 2.S3C**).

$$c(t) = c_{1/2} - \frac{1}{k}\log(\frac{F_m}{\Delta F/F_{Synth}} - 1), \tag{2.3}$$

$$r(t) = \frac{1}{\Delta t} T(\tau_r, \tau_d, t)^{-1} \bigotimes c(t) \tag{2.4}$$

where the parameters are a rise-time, τ_r; a decay-time, τ_d; a non-linearity parameter, k, NL; a half-activation parameter, $c_{1/2}$, EC50; and a maximum possible fluorescence change, F_m. $\bigotimes$ presents the operation of convolution between the inverse of spike-to-calcium kernel, $T(\tau_r, \tau_d, t)$, and the intermediate calcium variable, $c(t)$, where

$$T(\tau_r, \tau_d, t) = \exp(-\frac{t}{\tau_d})[1 - exp(-\frac{t}{\tau_r})].$$

For nonlinear decoders, we performed the direct inverse function based on the values of parameters that generated noisy $\Delta F/F_{Synth}$ dynamics. Since the external noise was strongly amplified near the saturation of nonlinearity, when applying the inverse function, we truncated the data to within 2% to 98% of the maximum fluorescence

change. That was

$$\Delta F/F_{Synth} = \begin{cases} 0.02F_m, & \text{if } F/F_{Synth} \leq 0.02F_m \\ 0.98F_m, & \text{if } F/F_{Synth} \geq 0.98F_m \end{cases}.$$

2. C2S Random linear decoder

To obtain distributions of decoded data, we performed the direct inverse function based on values of a set of parameters randomly chosen from those that could generate noisy $\Delta F/F_{Synth}$ dynamics. For random nonlinear decoder, we applied inverse functions using **Equations 2.3**, and **2.4**; for random linear decoder, we only performed the deconvolution based on **Equations 2.4** (where τ_r and τ_d were both randomly samples from their distributions).

3. C2S MCMC decoder

The spike inference algorithm was an adapted version of original work from (Pnevmatikakis et al., 2014a). This is an extension of their previously developed constrained deconvolution method (Pnevmatikakis et al., 2016), which incorporates a Monte Carlo Markov Chain (MCMC) sampler to explore underlying spike trains with super-resolution in time (Vogelstein et al., 2010; Vogelstein et al., 2009). In our version of code, we first performed the constrained deconvolution method (Pnevmatikakis et al., 2016) as the initialization for MCMC sampler (Pnevmatikakis et

al., 2014a), and if the MCMC improved little compared to the constrained deconvolution method (i.e., < explained variance of data in the constrained deconvolution method), we adopted results from the constrained deconvolution method; otherwise, we adopted the results from MCMC sampler. Details of the algorithm were described in the papers accordingly. In our hands, the MCMC decoder outperformed three other unsupervised decoders (Vogelstein et al., 2010; Vogelstein et al., 2009; Pnevmatikakis et al., 2016) in our comparison of performance of decoding spikes from raw imaging dynamics using simultaneous imaging-ephys datasets.

2.5.2 Supplementary figure legends

Figure 2.S1: Supporting figure for Figure 2.2: details in spike-to-calcium models.

(**A**) Distributions and pairwise correlations of parameters in S2C models. We collected the ensembles of cells for 5 parameters (i.e. a rise-time, τ_r; a decay-time, τ_d; a non-linearity parameter, NL; a half-activation parameter, EC50; and a maximum possible fluorescence change, F_m) in different imaging conditions (GCaMP6f viral delivery, 6f-AAV, 11 cells, 37 sessions, gray; GCaMP6s viral delivery, 6s-AAV, 9 cells, 21 sessions, yellow; GCaMP6f transgenic mouse, GP5.17, 18 cells, 32 sessions, purple; GCaMP6s transgenic mouse, GP4.3, 22 cells, 33 sessions, green), using simultaneous loose-seal electrophysiological recordings and imaging (see **Table 2.S2** for details of

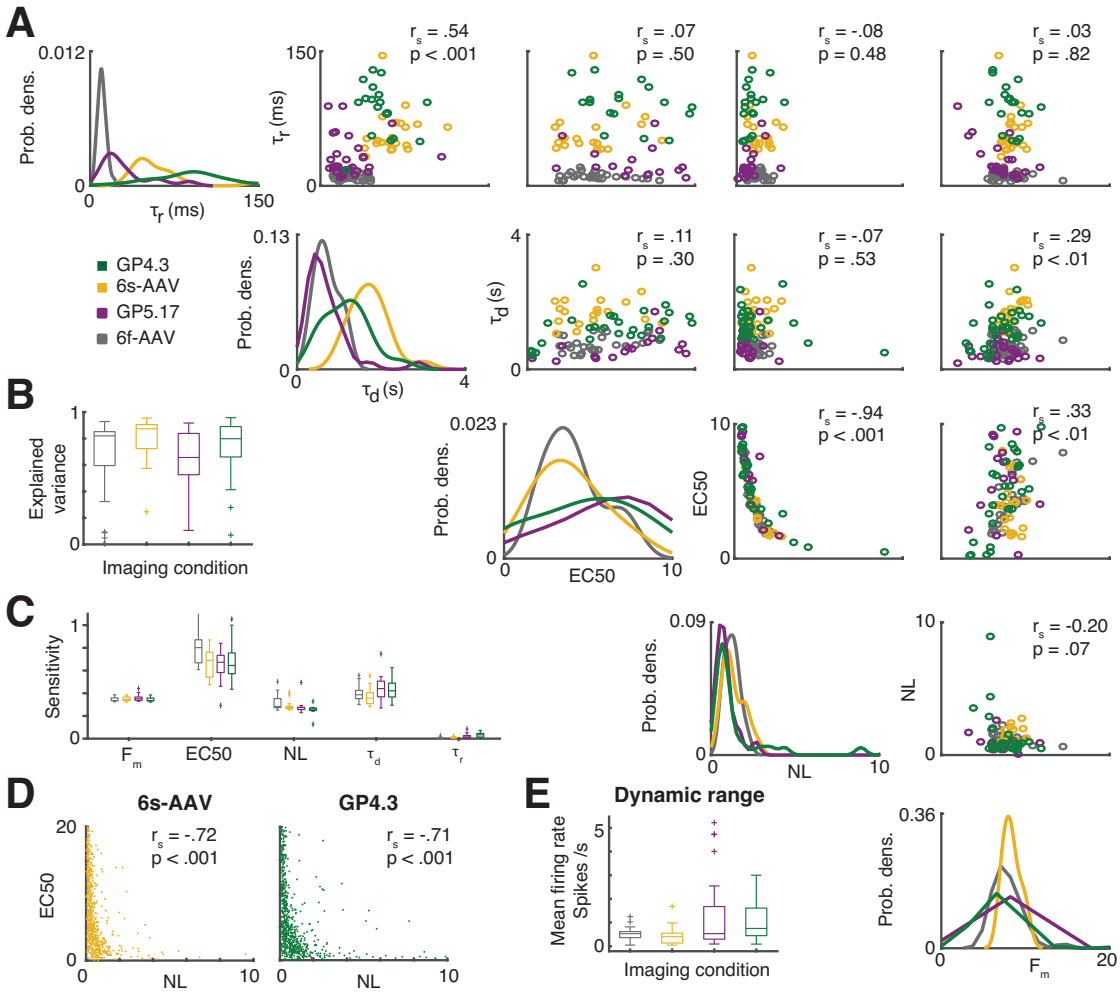


Figure 2.S1: Supporting figure for Figure 2.2: details in spike-to-calcium models.

datasets). Panels along the diagonal describes the distribution of each parameter; panels in off-diagonal describes the correlation between two parameters. Spearman correlation of parameters across cells (regardless of its recording methods) and its p-value were provided in each off-diagonal panel. Of note, the correlations were weak except that of NL and EC50 parameters. The strong inverse correlation between NL and EC50 resulted in the transition of nonlinear from sublinear to linear occurring at similar thresholds. Overall, the 6s-AAV data had slower dynamics than the GP4.3 data (rank-sum test, $p = .02$) (Dana et al., 2014), while the NL was similar (rank-sum test, $p = .40$). (**B**) Box-plots of explained variance of S2C on validation data for simultaneously recorded neurons (color follows the same convention as that in **Figure 2.S1A**). We computed first $\Delta F/F_{Synth}$ from spike train and estimated parameters in **Figures 2.S1A** (**Materials and Methods**, **Equations 2.1 and 2.2**) and then computed explained variance, R^2, as

$$R^2 = 1 - \frac{(\Delta F/F_{Data} - \Delta F/F_{Synth})^2}{(\Delta F/F_{Data} - < \Delta F/F_{Data} >)^2}.$$

The model typically fitted $R^2 > .6$ of the explained variance in the raw fluorescence dynamics. Given the criterion, 27/37 6f-AAV neurons ($R^2 = .82 \pm .27$; median std.), 19/21 6s-AAV neurons ($R^2 = .87 \pm .17$), 19/32 GP5.17 neurons ($R^2 = .66 \pm .23$), 28/33 GP4.3 neurons ($R2 = .80 \pm .20$) were passed. (**C**) Box-plot of distribution of parameter sensitivity values. The distribution of each parameter spans a wide

range across cells. If the fit of a parameter were insensitive to the data (i.e., a sloppy parameter), a range of values of this parameter would result in the same quality of fit, and therefore a wide range of the estimated distribution could stem simply from random end points of the fit across neurons. To confirm that such a wide range did not result from sloppiness in the fits, we performed sensitivity tests independently for each parameter. Sensitivity was defined as the decrease of the fraction of explained variance, as a function of the deviation of the parameter value from the estimated solution,

$$g = \frac{\Delta EV/EV}{\Delta P/P},$$

where $P \in \{\tau_r, \tau_d, NL, EC50, F_m\}$. We found that all the parameters, except τ_r (see **Table 2.S4** for details of datasets), had considerable sensitivities, i.e. a one-fold change in the parameter reduced the explained variance by approximately 40%, implying that the wide range of parameter value we found reflects variability among cells. (**D**) Pairwise correlation of refits of NL and EC50 using ALM imaging dynamics. We performed a refit of parameters in modeling ALM synthetic imaging dynamics, based on non-simultaneous data, by estimating, for each neuron's NL parameters that would have led to the closet fit with the best possible partner across all imaged cells (each target neuron was used less than 5 times to avoid biases). Since the single spike parameters were tested well within the data, we kept the relevant parameters fixed and refit only NL parameters. Distributions of refits of NL parameters strongly overlapped with those obtained in simultaneous imaging-ephys recordings, and NL

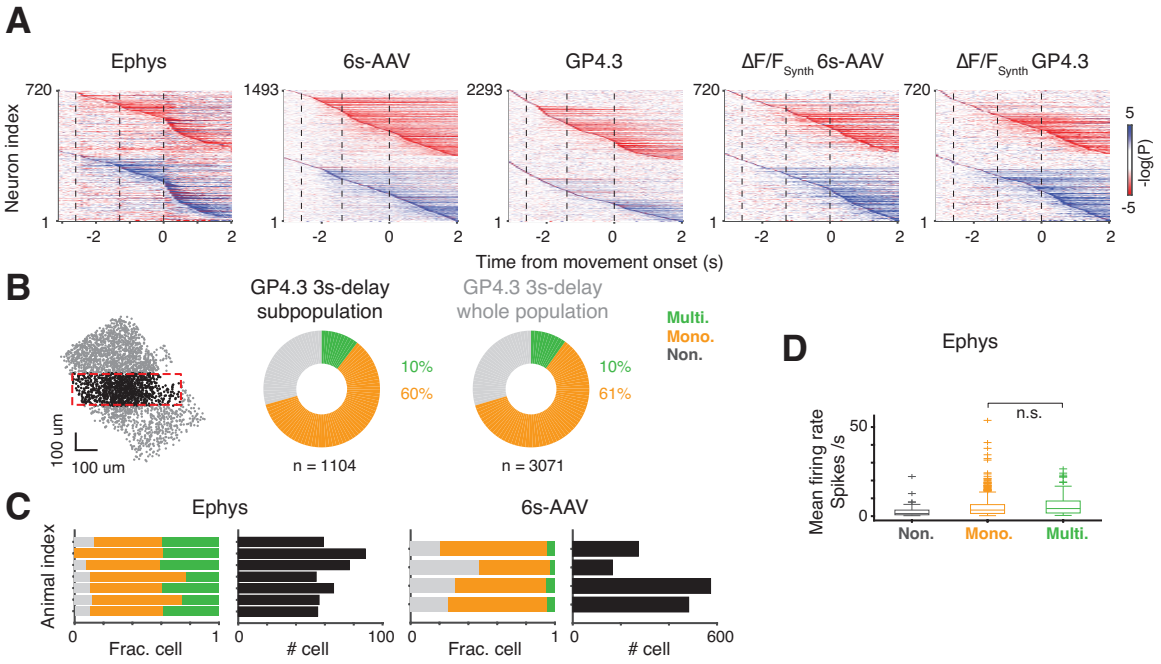


Figure 2.S2: Supporting figure for Figure 2.3: dynamics of selectivity in different imaging conditions.

and its midpoint similarly had a similar strong inverse correlation ($r_s < -.7$, $p < .001$). (**E**) Box-plots of firing rates of cells in each recording sessions (6f-AAV, gray, 0.51 ± 0.25 Hz, mean±std., range 0.05—1.25 Hz; 6s-AAV, yellow, 0.43 ± 0.38 Hz, range 0.05—1.68 Hz; GP5.17, purple, 1.25 ± 1.48 Hz, range 0.09—5.22 Hz; GP4.3, green, 1.08 ± 0.85 Hz, range 0.09—3.00 Hz), which were in between zero —6 Hz.

Figure 2.S2: Supporting figure for Figure 2.3: dynamics of selectivity in different imaging conditions.

(**A**) Dynamics of selectivity. To quantify the change of instantaneous selectivity (from left to right, Ephys., 6s-AAV, GP4.3, S2C 6s-AAV, and S2C GP4.3), we com-

puted the p-values of neural preference using two-sample t-test. -log(p) value of each neuron is shown as a function of time in each panel, in which white area of color bar ranges from -3 to 3 (where the p-value is $> .05$) and the other area indicates a preference for lick right (trial type A, blue) or lick left (trial type B, red). (**B**) Fractions of mono- (orange), multiphasic (green) and nonselective (gray) neurons were robust to the choice of anterior-to-posterior and medial-to-lateral locations of recording, based on the imaging data from GP4.3 with 3 s delay (χ^2 test, $p = .52$). Black area, the recording area in imaging was similar to that in ephys, n = 1104 cells; gray area, all the recording area in imaging, n = 3071 cells. (**C**) Fractions of mono- (orange), multiphasic (green) and nonselective (gray) were robust to the choice of animals of recording. 7 animals in ephys with numbers of neuron > 50; 4 animals in 6s-AAV imaging with numbers of neuron > 50. χ^2 test, ephys: $p = .07$; 6s-AAV: $p = .81$. (**D**) Firing rates of ephys. neurons in three categories. Extracellular electrophysiology tended to sample neurons firing at high rates. Here we examined whether mono- and multiphasic neurons have different baseline firing rate. Box-plots of mean firing rate distributions of the non-selective, mono-, and multiphasic neurons (from left to right). Across population, the mean firing rates were similar in mono- to that in multiphasic neurons (two-sample t-test, two-tail, $p = .46$). Firing rates: non-selective neuron, 2.42 ± 3.07 Hz (mean±std.; n = 79), range 0.26—22.24 Hz; monophasic selective neuron, 5.46 ± 6.25 Hz (mean±std.; n = 421), range 0.27—53.74 Hz; multiphasic selective neuron, 5.82 ± 5.22 Hz (mean±std.; n = 220), range 0.40—26.54 Hz.

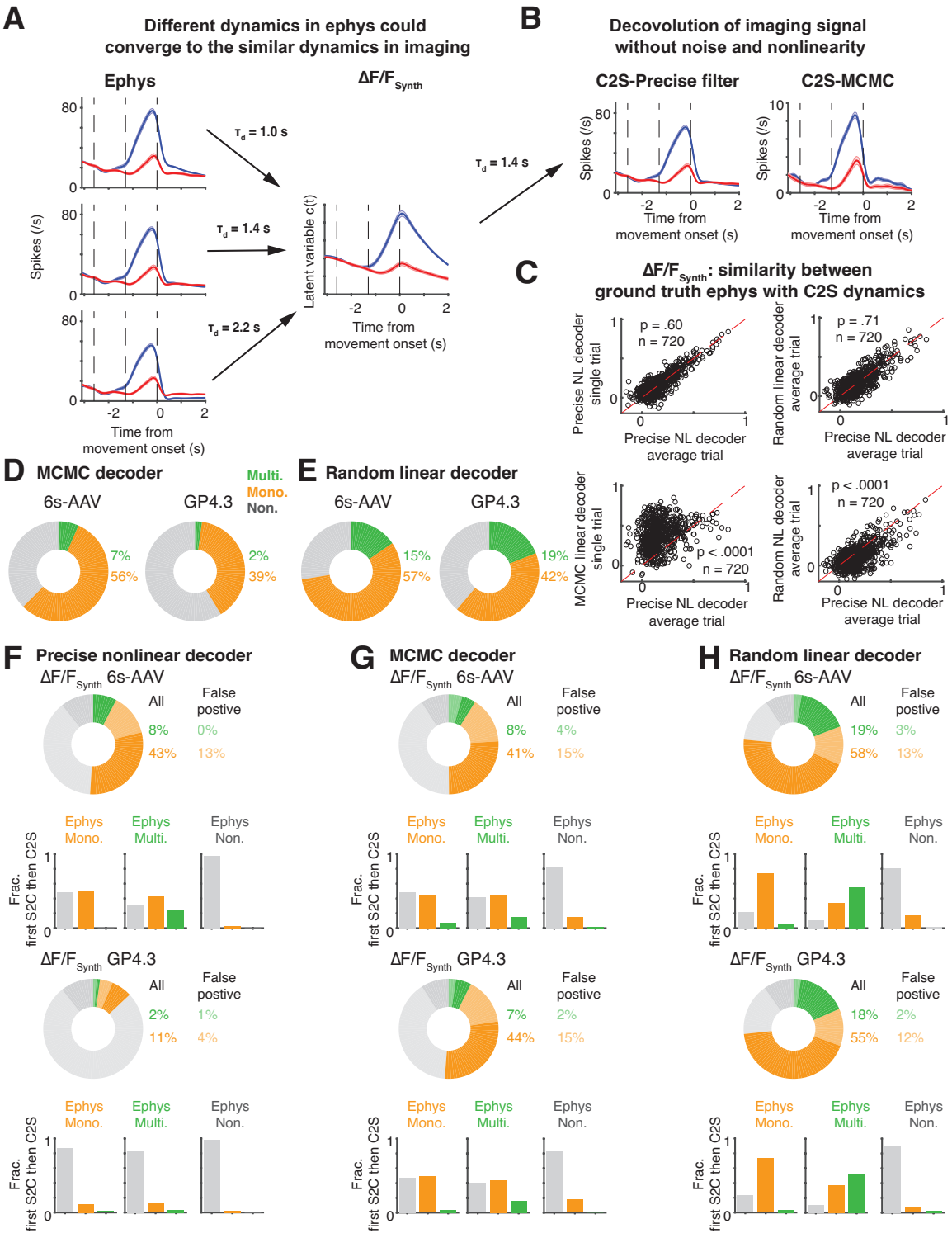


Figure 2.S3: Supporting figure for Figures 2.2-2.3: details in calcium-to-spike (C2S) models and inferred ephys from imaging using C2S model can account a fraction of multiphasic neuron in imaging.

Figure 2.S3: Supporting figure for Figures 2.2-2.3: details in calcium-to-spike (C2S) models and inferred ephys from imaging using C2S model can account a fraction of multiphasic neuron in imaging.

(**A**) Schematic description that different dynamics in ephys could lead to the same dynamics in imaging. We took three neurons in ephys data (from top to bottom, mono-, mono-, and multiphasic selective neurons), where they have the similar dominant dynamics (divergence of dynamics from sample to delay), while a salient difference in the secondary dynamics (dynamics in response). We found that their $\Delta F/F_{Synth}$ dynamics could converge (since $c(t)$ is identical) while being convolved with different values of decay time, τ_d (from left to right). (**B**) Schematic description of difference between precise decoder and MCMC decoder. Left: precise decoder can recover precisely the mean dynamics of ephys from $c(t)$ without internal noise. Right: MCMC decoder seems to converge to the one of the possible ephys dynamics, that captured the dominant mean dynamics of ephys from $c(t)$, but failed to recover the secondary dynamics. (**C**) Comparison of performance across different decoders. Here we compared performance (defined as the similarity index, Spearman's correlations, between averaged dynamics of ground truth ephys that generated noisy $\Delta F/F_{Synth}$ and that inferred from $\Delta F/F_{Synth}$) using several combination of the direct decoders and the MCMC decoder. Left top: precise nonlinear decoder on single trials vs pre-

cise nonlinear decoder on average across trials (identical, paired t-test, $p = .60$, n = 720); Right top: random linear decoder on average across trials vs precise nonlinear decoder on average across trials (identical, paired t-test, $p = .71$, n = 720); Left bottom: MCMC decoder on single trials vs precise nonlinear decoder on average across trials (MCMC is better, paired t-test, $p < .0001$, n = 720); Right bottom: random nonlinear decoder on average across trials vs precise nonlinear decoder on average across trials (precise nonlinear decoder is better, paired t-test, $p < .0001$, n = 720). We therefore applied (1) three decoders (i.e. precise nonlinear decoder using single trial; MCMC using single trial; random linear decoder using average across trials) in comparison of recoverability of multiphasic dynamics (**Figures 2.S3F, G, H**, respectively) and peakiness of dynamics (**Figures 2.S4E, F, G**, respectively) in synthetic imaging $\Delta F/F_{Synth}$ dynamics and (2) two decoders (i.e. MCMC using single trial; random linear decoder using average across trials) in comparison of recoverability of multiphasic dynamics (**Figures 2.S3D, E**, respectively) and peakiness of dynamics (**Figures 2.S4C, D**, respectively) in imaging dynamics. (**D**) Fractions of mono- (orange), multiphasic (green) and nonselective (gray) in inferred ephys from 6s-AAV (left) and GP4.3 (right) imaging using MCMC decoder on single trial. (**E**) The same plot as **D** using C2S Random linear decoder on averaged dynamics across trials. To evaluate the significant of selectivity across trials, we modeled the variability (vari-

ance,) as a linear function of inferred firing rate, $r(t)$.

$$\sigma_r(t) = \sqrt{\alpha \cdot r(t)}$$

where $\alpha = 0.6$ is the Fano factor. (**F**) Fractions of neurons classified as mono- (orange), multiphasic (green) and nonselective (gray) in inferred ephys from S2C 6s-AAV (top) and GP4.3 imaging (bottom) using precise nonlinear decoder on single trials; transparent (light color) region presents the fraction of inferred ephys neurons that were mislabeled when compared to ground truth ephys. Since the ground truth of each cell type was known in ephys, we computed the fraction of monophasic- (left panels of bar plot), multiphasic- (middle panels of bar plot), and non-selective (right panels of bar plot) neurons that became monophasic- (orange bar), multiphasic- (green bar), and non-selective (gray bar) neurons after precise nonlinear decoder using synthetic imaging data. We also estimated the fraction of neuron being mislabeled after precise nonlinear decoder using synthetic imaging data (False positive rate; second columns on the right of the pie charts). (**G**) The same plot as **F** using C2S MCMC decoder on single trials. (**H**) The same plot as **F** using C2S Random linear decoder on averaged dynamics across trials, where we modeled variability using $\alpha = 2.5$. In general, we found that random linear decoder gave a larger recovery of multiphasic neurons in both imaging data and synthetic imaging data (**Figures 2.S3E, H**, respectively).

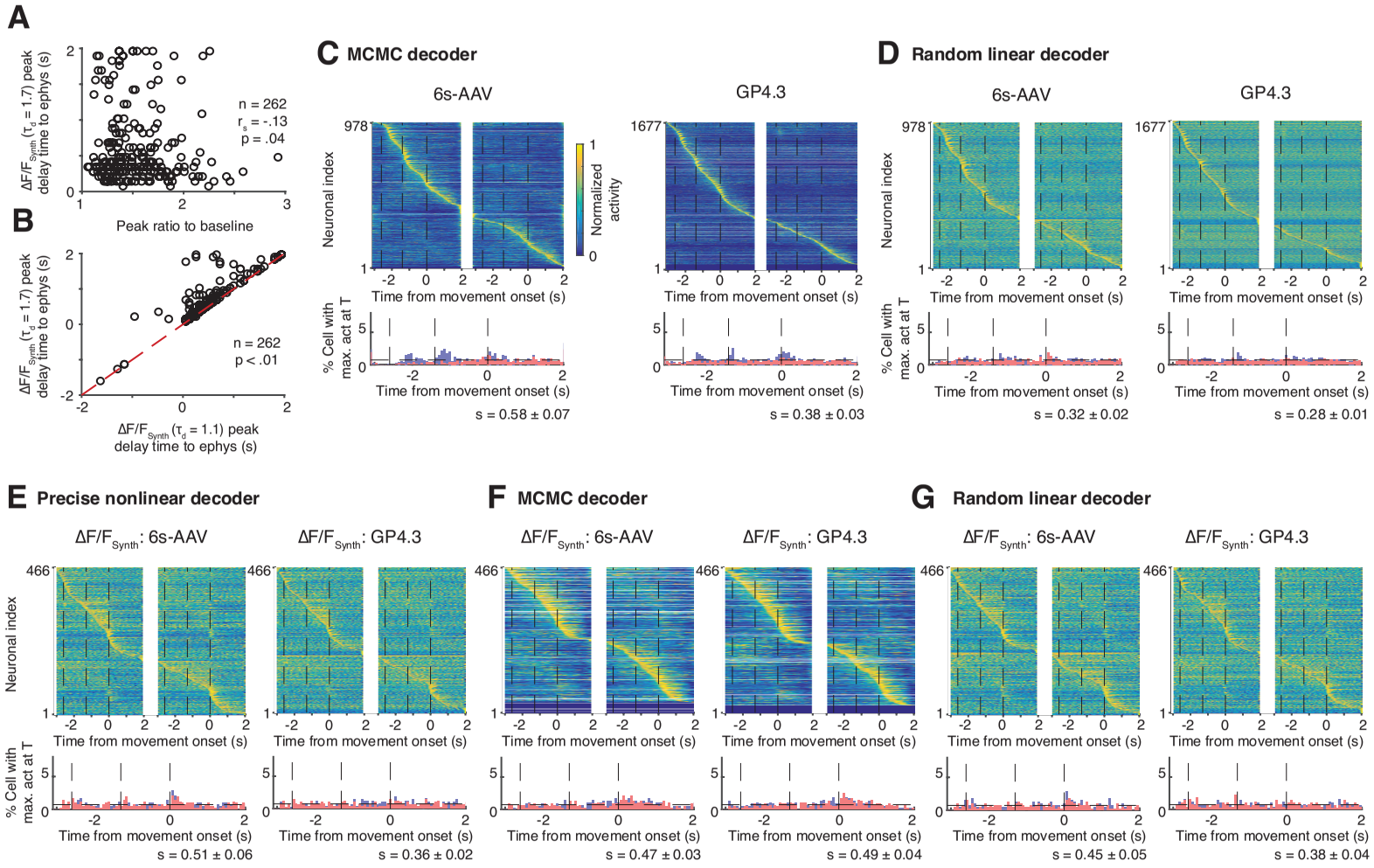

A
B
C MCMC decoder
D Random linear decoder
6s-AAV
GP4.3
Neuronal index
Normalized activity
Time from movement onset (s)
% Cell with max. act at T
s = 0.58 ± 0.07
s = 0.38 ± 0.03
s = 0.32 ± 0.02
s = 0.28 ± 0.01
n = 262
r_s = -.13
p = .04
p < .01
Peak ratio to baseline
ΔF/F_Synth (τ_d = 1.7) peak delay time to ephys (s)
ΔF/F_Synth (τ_d = 1.1) peak delay time to ephys (s)
E Precise nonlinear decoder
F MCMC decoder
G Random linear decoder
ΔF/F_Synth: 6s-AAV
ΔF/F_Synth: GP4.3
s = 0.51 ± 0.06
s = 0.36 ± 0.02
s = 0.47 ± 0.03
s = 0.49 ± 0.04
s = 0.45 ± 0.05
s = 0.38 ± 0.04

Figure 2.S4: Supporting figure for Figure 2.5: inferred ephys from imaging using calcium-to-spike model can account for little time-lock dynamics.

(**A**) Delay time of peak in synthetic imaging dynamics to that in ephys decreases as a function of baseline activity in ephys ($r_s = .13$, $p = .04$). We computed the delay time of peak in synthetic imaging dynamics to that in ephys using the neurons with peaks at time zero (onset of response; n = 262). (**B**) Delay time of peak in synthetic imaging dynamics to that in ephys decreases as a function of decay time (x-axis, $\tau_d = 1.1$ s; y-axis, $\tau_d = 1.7$ s; n = 262; paired t-test, $p < .01$; red line, diagonal line). (**C**) Top: Heatmap of normalized trial-averaged firing rates for lick left trials (left column) and lick right trials (right column) for inferred ephys data from 6s-AAV (left) and GP4.3 (right) imaging data using MCMC decoder on single trials. Firing rates were normalized to the maximum of activity across both conditions and neurons were sorted by latency of peak activity and by their preferred trial type. Bottom: Fraction of neurons with a peak at given time point over time. distribution in time (inferred ephys from 6s-AAV, left; that from GP4.3, right) plotted simultaneously for both trial types (red: right preferring trials, blue: left preferring trials, black horizontal line: uniform distribution). The peakiness level ($s = 0.58 \pm 0.07$, 6s-AAV; $s = 0.38 \pm 0.03$,

Figure 2.S4 *(preceding page)*: Supporting figure for Figure 2.5: inferred ephys from imaging using calcium-to-spike model can account for little time-lock dynamics.

GP4.3) remains low comparing to that of ephys (both $p < .001$, bootstrap, t-test). (**D**) The same plot as **C** using C2S Random linear decoder on averaged dynamics across trials. The peakiness level ($s = 0.32 \pm 0.02$, 6s-AAV; $s = 0.28 \pm 0.01$, GP4.3) remains low comparing to that of ephys (both $p < .001$, bootstrap, t-test). (**E**) The same convention of plot as **C**, but using the synthetic imaging data (S2C 6s-AAV, left; S2C GP4.3, right) and using precise nonlinear decoder on single trials. The peakiness level ($s = 0.51 \pm 0.06$, 6s-AAV; $s = 0.36 \pm 0.02$, GP4.3) remains low comparing to that of ground truth ephys (both $p < .001$, bootstrap, t-test). (**F**) The same as **E** using C2S MCMC decoder on single trials. The peakiness level ($s = 0.47 \pm 0.03$, 6s-AAV; $s = 0.49 \pm 0.04$, GP4.3) remains low comparing to that of ground truth ephys (both $p < .001$, bootstrap, t-test). (**G**) The same as **E** using C2S Random linear decoder on averaged dynamics across trials. The peakiness level ($s = 0.45 \pm 0.05$, 6s-AAV; $s = 0.38 \pm 0.04$, GP4.3) remains low comparing to that of ground truth ephys (both $p < .001$, bootstrap, t-test). In summary, we found that on our hand, both direct deconvolution and MCMC inference model would fail to undo the time-lock activity pattern from the sequence-like dynamics in both imaging and synthetic imaging data.

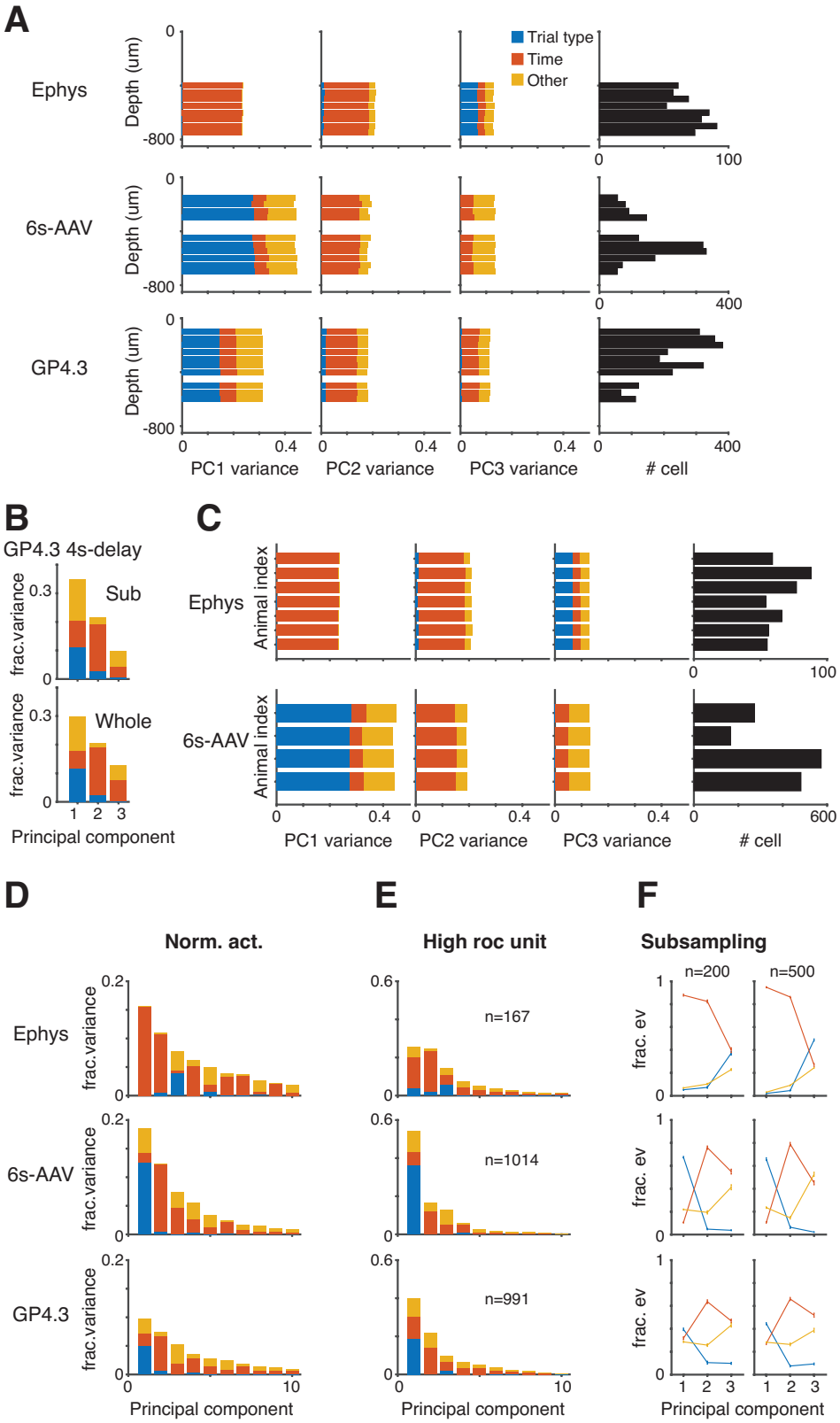

A
Ephys
6s-AAV
GP4.3
Depth (um)
0
-800
Trial type
Time
Other
0
100
400
0.4
PC1 variance
PC2 variance
PC3 variance
cell
B
GP4.3 4s-delay
Sub
Whole
frac.variance
0.3
1 2 3
Principal component
C
Animal index
600
D
Norm. act.
0.2
10
E
High roc unit
0.6
n=167
n=1014
n=991
F
Subsampling
n=200
n=500
frac. ev
1

Figure 2.S5: Supporting figure for Figure 2.6: explained variance of first three PCs are robust to the specificity of the confounding factors in comparison.

In one of our main results, we found that the fraction of explained variance (EV) due to temporal dynamics was high in the 1st PC in the ephys dataset ($98.71 \pm 0.06\%$, mean std, bootstrap), whereas trial-type selectivity was high in the 1st PC of imaging (6s-AAV: $60.39 \pm 0.29\%$; GP4.3: $44.51 \pm 0.65\%$) (**Figures 2.6A-C**). To control for potential confounding factors, we confirmed that this observation was robust to (1) location of sampling of subpopulation of neurons (depth: **Figures 2.S5A**; area, **Figures 2.S5B**); (2) animal identity (**Figures 2.S5C**); (3) normalized or raw neural activity (**Figures 2.S5D**); (4) thresholds of selectivity index (**Figures 2.S5E**); (5) size of subpopulation in analysis (**Figures 2.S5F**). Particularly, we found that, in **Figures 2.S5D-F**, (1) EV of 1st PC was high in time in ephys and that is high in trial in imaging; (2) EV of 2nd PC was high in time, were consistent in first 2 PCs regardless normalization, selectivity or subsample size of the data. EV contents were, however, significantly different in analyses for higher-order PCs. For example, in ephys, EVs of time and trial type in 3rd PC were similar at small subsample ($p > .05$, paired t test), while that is significantly high in trial type at high subsample

Figure 2.S5 *(preceding page)*: Supporting figure for Figure 2.6: explained variance of first three PCs are robust to the specificity of the confounding factors in comparison.

($p < .001$, paired t test).

(**A**) EV of first three PCs (time, red; trial type, blue; other, yellow) was robust to the choice of depth of recording (the same convention as **Figure 2.3G**; $p > .05$, anova; 1000 subsampling; for each subsample, we randomly sampled 20 trials from each condition, i.e. trial type A and B). (**B**) EV of first three PCs was robust to the choice of anterior-to-posterior and medial-to-lateral locations of recording (the same convention as **Figure 2.S2B**; $p > .05$, anova; 1000 subsampling). (**C**). EV of first three PCs was robust to the choice of animals of recording (the same convention as **Figure 2.S2C**; $p > .05$, anova; 1000 subsampling). (**D**) EV of 1st PC was high in time in ephys and that is high in trial in imaging ($p < .001$, paired t test, 1000 subsampling; pairs of EVs of time and trial type in each subsampling). EV of 2nd PC was high in time ($p < .001$, paired t test, 1000 subsampling; pairs of EVs of time and trial type in each subsampling). (**E**) The same plot as **D**, using high selective neuron ($roc > .55$ for all three epochs, i.e. sample, delay, and response; $p < .001$ for all paired t test). (**F**) The same plot as **D**, using random subsamples of 200 (right) and 500 (left) units ($p < .001$ for all paired t test).

2.5.3 Supplementary tables

Table 2.S1 and 2.S2: Datasets details.

Table 2.S3 and 2.S4: Spike-to-calcium model parameter summaries.

[1]Public source: http://crcns.org/data-sets/motor-cortex/alm-1
[2]Public source: http://crcns.org/data-sets/motor-cortex/alm-2
[3]Public source: http://crcns.org/data-sets/methods/cai-1
[4]high-zoom imaging condition
[5]Public source: http://crcns.org/data-sets/methods/cai-1
[6]median ± std.
[7]high-zoom imaging condition
[8]mean ± std.
[9]high-zoom imaging condition

Table 2.S1: Delayed discrimination task

Dataset name	Delay time (s)	# cells	# animals	Reference
Ephys[1]	1.3	720	19	Li et al., Nature 2015; Nuo Li, Charles R Gerfen, Karel Svoboda (2014)
6s-AAV[2]	1.4	1493	4	Li et al., Nature 2015; Tsai-Wen Chen, Nuo Li, Charles R Gerfen, Zengcai V. Guo, Karel Svoboda (2016)
GP4.3	1.4	2293	1	Dana et al., Plos One, 2014
GP4.3	3.0	3071	2	Chen et al. (to be published)

Table 2.S2: Simultaneous ephys-imaging experiments

Dataset name	# cells	# sessions	Reference
6s-AAV[3]	9	21	Chen, et al., (2013); GENIE project, Janelia Farm Campus, HHMI; Karel Svoboda (contact). (2015)
GP4.3[4]	22	33	
6f-AAV[5]	11	37	Chen, et al., (2013); GENIE project, Janelia Farm Campus, HHMI; Karel Svoboda (contact). (2015)
GP5.17	18	32	

Table 2.S3: Spike-to-calcium model parameter range

Imaging condition	τ_r (ms)	τ_d (s)	NL	EC50	F_m
6s-AAV	50.54 ± 24.56[6]	1.71 ± 0.46	1.13 ± 0.65	3.97 ± 1.91	0.82 ± 0.10
GP4.3[7]	92.70 ± 37.51	1.23 ± 0.54	0.89 ± 5.00	5.30 ± 3.46	0.71 ± 5.74
6f-AAV	10.23 ± 3.69	0.68 ± 0.27	1.24 ± 0.44	3.87 ± 1.76	0.75 ± 1.79
GP5.17	21.23 ± 22.25	0.57 ± 2.48	0.77 ± 5.08	7.17 ± 24.71	0.77 ± 18.14

Table 2.S4: Spike-to-calcium model parameter sensitivity

Imaging condition	τ_r	τ_d	NL	EC50	F_m
6s-AAV	0.01 ± 0.01[8]	0.37 ± 0.08	0.29 ± 0.04	0.67 ± 0.13	0.35 ± 0.02
GP4.3[9]	0.03 ± 0.02	0.43 ± 0.08	0.26 ± 0.04	0.68 ± 0.15	0.34 ± 0.02
6f-AAV	0.01 ± 0.01	0.39 ± 0.06	0.32 ± 0.07	0.82 ± 0.16	0.34 ± 0.02
GP5.17	0.02 ± 0.02	0.44 ± 0.12	0.27 ± 0.06	0.66 ± 0.18	0.35 ± 0.03

Chapter 3

Single-trial dynamics of premotor cortex predicts behavioral variability

Neurons in the anterolateral motor cortex (ALM; premotor area) exhibit rich movement-related dynamics. To connect the neural dynamics at different stages of a decision-making task, we developed a time-varying linear dynamics system model that uncovered the shared activities among neurons. We found that the shared neuronal activity supported the continuous dynamics of the choice generation and memory in a fashion akin to drift diffusion models, and found a robust persistent decision signal in the post-decision period, which *de facto* associated the pre-sample activity with a previous decision. Importantly, we identified that error trials followed similar dy-

namics to those of correct trials, but their dynamics were separated in shared-activity space, providing an accurate early estimation of reward. Overall, the neural dynamics in shared-activity space could predict multiple measures of behavioral variability including performance, reaction time, and correctness, and therefore are a useful summary of the neural representations. Such an approach can be readily applied to study complex dynamics in other neural systems.

3.1 Introduction

Decision-making is a key component in the study of cognition (Gold and Shadlen, 2007; Shadlen and Shohamy, 2016; Shadlen and Kiani, 2013). In such paradigms, an animal integrates sensory inputs, generates a tentative behavioral choice, stores it in memory, executes it and finally evaluates its choice compared to the expected delivery of reward. This process involves multiple brain areas (Gold and Shadlen, 2007; Shadlen and Newsome, 2001; Shadlen and Shohamy, 2016; Shadlen and Kiani, 2013; Kepecs et al., 2008; Romo et al., 1999; Guo et al., 2014b; Guo et al., 2015; Brody et al., 2003; Stuphorn and Schall, 2006; Stuphorn et al., 2010). A certain degree of specialization exists among these neural circuits with respect to the different decision components. Nevertheless, activity in frontal brain areas holds a mixed representation of sensory, choice and reward signals (Brody et al., 2003; Romo et al., 1999; Rigotti et al., 2013). Moreover, such areas often have strong dynamics in

which the representation may potentially transform from one decision component to another. For example, electrophysiological recordings reveal complicated dynamics of ALM neurons associated with an animal's internal state, to prepare for movement in decisions (Guo et al., 2014b; Li et al., 2015; Li et al., 2016). However, decision neurons were untangled with those in post-decision activity at the onset of response time, which resulted in a discontinuity of information flow in neural space (Li et al., 2016). We therefore examined whether a local neural circuit (e.g. that of tens of neurons) in ALM could provide a continuous neural signal underlying the internal states in different stages of a decision. We identified such a signal in a latent space, where each neural mode presents a source of input shared by multiple neurons. Moreover, we found that neural-mode dynamics in this shared-activity space could predict the behavioral variability, such as trial type, reaction time and correctness, in single trials, and be robustly maintained for seconds in post-decision, representing single-trial identities.

As neuroscience focuses more on population recordings and single-trial analyses, an important question is how accurately behavioral variability can be decoded from neural activity (Cunningham and Yu, 2014). An important component in this question is how long useful decodable information lasts. Or to put it differently, for a given level of decodability of behavioral variability from immediately preceding neural activity, what is the level of decodability as one moves back through time. If decodability remains high and consistent, then models that incorporate this past in-

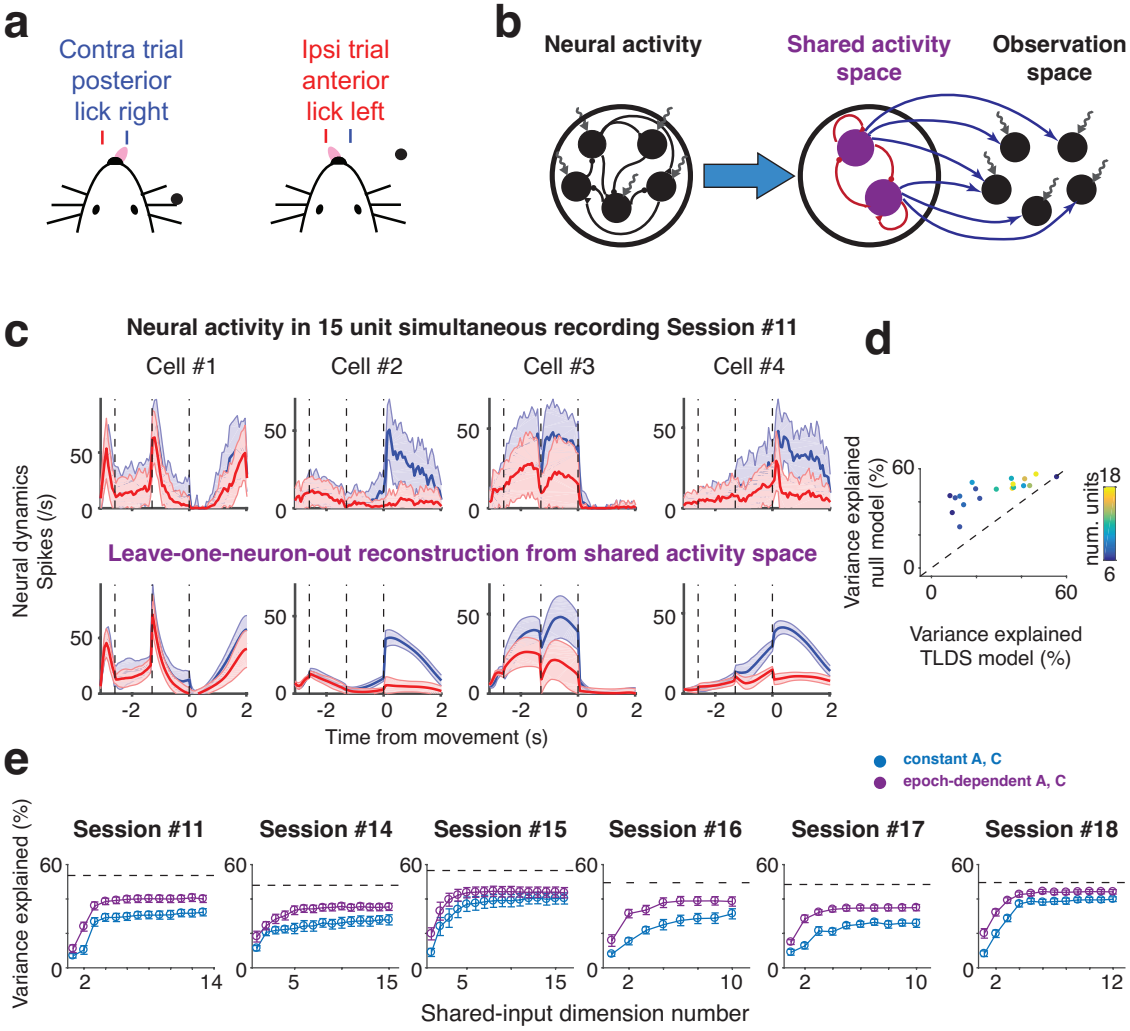
formation can improve in accuracy. The complex dynamics of activity in frontal lobe activity may suggest that information gets quickly wiped out, however, this is not true for all types of dynamics.

3.2 Results

3.2.1 Neural dynamics and the shared-activity space

Mice performed a delayed discrimination task of pole location, while multi-unit recordings were obtained from left ALM. After a brief delay (1.3 s), an auditory 'go' cue signaled the onset of response, and mice reported their choices by licking one of two ports (posterior to lick right, contra-trial; anterior to lick left, ipsi-trial; **Figure 3.1a**). ALM serves a significant role in planning directional licking movements (Komiyama et al., 2010; Guo et al., 2014b; Li et al., 2015). Indeed, most recorded ALM neurons ($n = 1,563$; 26 mice) were selective to some aspect of the task ($n =$ 1,296; **Figure 3.S1a**).

To perform single-trial analysis and relate activity to behavior, it would be advantageous to use population decoding methods instead of relating the activity of each single neuron to behavioral variability. Once a time-bin for such analysis has been chosen, standard statistical techniques can be used for performing decoding. However, the decision regarding which time scale to use is not trivial. On the one hand, a behaviorally relevant state is likely to persist for a substantial amount of time and


a
Contra trial
posterior
lick right
Ipsi trial
anterior
lick left
b
Neural activity
Shared activity space
Observation space
c
Neural activity in 15 unit simultaneous recording Session #11
Cell #1
Cell #2
Cell #3
Cell #4
Neural dynamics
Spikes (/s)
Leave-one-neuron-out reconstruction from shared activity space
Time from movement (s)
d
Variance explained
null model (%)
Variance explained
TLDS model (%)
num. units
e
constant A, C
epoch-dependent A, C
Session #11
Session #14
Session #15
Session #16
Session #17
Session #18
Variance explained (%)
Shared-input dimension number

therefore one would prefer long time bins to improve sampling from low rate neurons. However, the readout of the neuronal state and its variability may change across time, and in particular across behavioral epochs (Churchland et al., 2010b; Churchland et al., 2011). Linking many independent short time bins is not likely to be successful since the number of variables to be estimated is equal to the number of neurons multiplied by the number of time points, so one is immediately limited by having fewer trials than variables to estimate.

Here we aim to resolve the tension between these goals by using a model based approach where we first estimate a dynamical latent space model (Roweis and Ghahramani, 1999; Ghahramani and Hinton, 1996a) that attempts to capture the evolution of shared activity across neurons, but also allow this model to change its parameters

Figure 3.1 *(preceding page)*: Neural activity of anterolateral motor cortex neurons in shared-activity space.
a. Schematic description of experiment. Mice were trained to report pole position by directional licking, interleaved by a brief delay (posterior, lick right, contra. trial; anterior, lick left, ipsi. trial). **b.** Schematic description of time-varying linear dynamical systems (TLDS) model. As the recorded neuronal activity shows strong correlation across time, the full-neural space of neural activity can be represented by a two-layer network (TLDS), where the latent network (purple circles) in shared-input space explicitly models the correlated dynamics of neurons (due to common inputs or direct recurrent connections) and the activity of each neuron is a projection up from the shared-input space combined with independent input. **c.** TLDS reproduces neural activity in simultaneous recordings using the other neural activities within the same pool. Top: Observed neuronal activity from 15-unit simultaneous recording session, bottom: prediction of neuronal activity on prediction data where the activity of that neuron is unobserved. Shaded area, *std.* across trials. **d**. Comparisons of variance in neural dynamics explained by TLDS fit and the mean activity model (null model). **e**. Comparisons of the model TLDS fit using different sets of the latent dimensions and **A, C** matrices (purple, epoch-dependent **A, C** matrices; blue, constant **A, C** matrices; black dash line, variance explained by the mean activity model).

across behavioral epochs to capture the well-known varying nature of dynamics across behavioral epochs (Petreska et al., 2011). We refer to this model as the time-varying linear dynamical system model (TLDS). The model is defined in the following fashion:

$$\mathbf{r}(t) = \mathbf{C}(s)\mathbf{x}(t) + \mathbf{r}_0 + \mathbf{v}(t) \tag{3.1}$$

$$\mathbf{x}(t) = \mathbf{A}(s)\mathbf{x}(t-1) + \mathbf{w}(t-1) \tag{3.2}$$

Whereby we assumed that the N-dimensional neural activity (full-neural space, FNS), $\mathbf{r}(t) \in \mathbf{R}^N$ (**Figure 3.1b**, left), can be well modeled by a low dimensional dynamical system and a projection that transforms this latent space into full activity. This can be thought of as a two-layer representation whereby the first layer consists of a small number, M ($M < N$), of neural modes: $\mathbf{x}(t) \in \mathbf{R}^M$**,** i.e., the shared activity space (**Figure 3.1b**, right; recurrent connection, epoch-dependent matrix $\mathbf{A}$, red lines; s, epoch index; latent inputs, $\mathbf{w} \sim N(0, \sigma^2_{int}(m))$; $\sigma_{int}(m)$, the amplitude inputs to mth neural mode); and a second layer that describes the projection from shared activity space (SAS) to the observed neuronal activity via a matrix $\mathbf{C}$, upon which neuron-independent input is added, $\mathbf{v} \sim N(0, \sigma^2_{ext}(n))$, $\sigma_{ext}(n)$, the amplitude of the independent input to nth neuron (**Figure 3.1b**, right; gray arrows); $\mathbf{r}_0 \equiv < \mathbf{r}(t) >_t$ is the mean activity across time and trials.

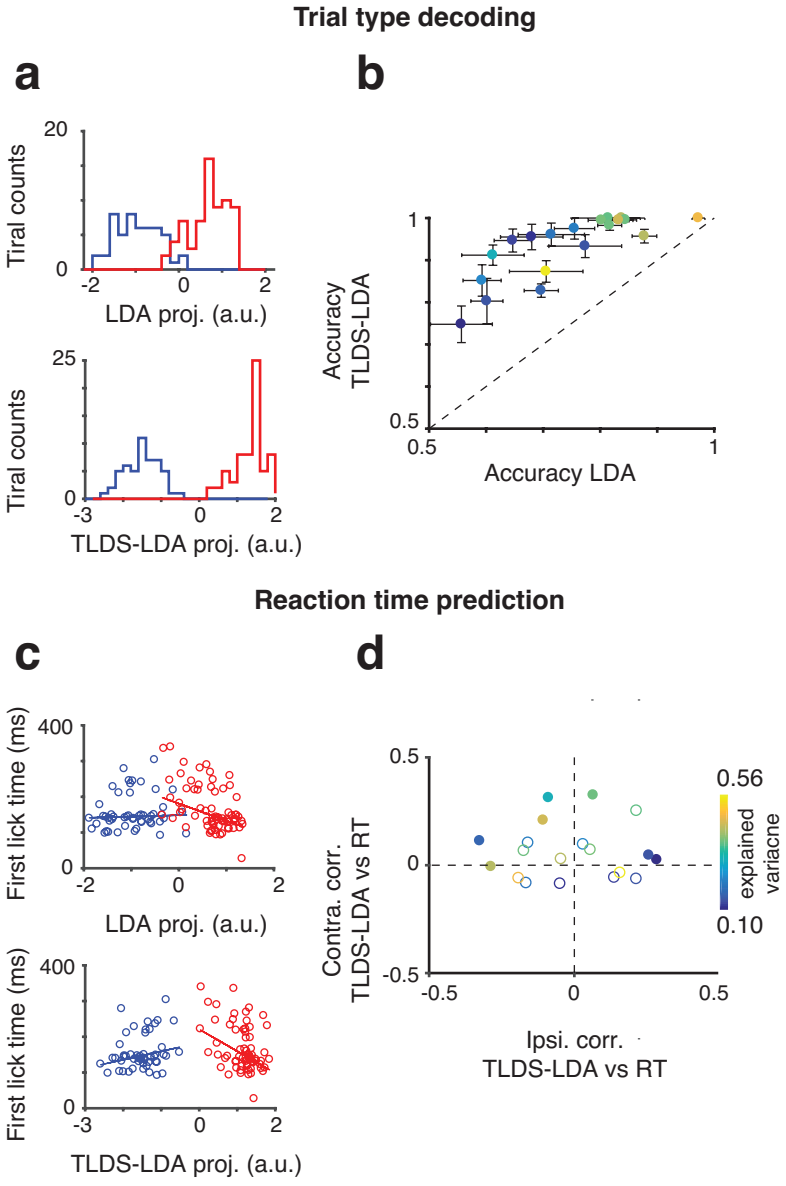
In order to measure the goodness of fit of the model when applying it to the data, we use an approach known as leave-one-neuron-out (LONO) (Yu et al., 2009). Intu-

itively, if the dynamics are composed of correlated modes, then once these dynamics have been estimated for the entire complement of neurons (training data) it should be possible to predict the activity of any given neuron from an estimation of shared activity space that is performed on test trials without including the activity of that particular neuron. Comparing the neural activity with its estimation, we determined the goodness of fit using TLDS and found that the neural modes in the SAS could explain quite well the activity of most neurons in this LONO scenario (**Figure 3.1c**) ($R^2 = .42$, **Figure 3.1c**; R^2 ranged from .09 to .56, 19 sessions, which increases with that from the average activities of neurons, $r_s = .80$, $p < .001$; **Figure 3.1d**).

Interestingly, we find that although the number of simultaneously recorded neurons spanned a fairly wide range, the dimensions of the SAS were similar (3 or 4; **Figure 3.1e**; **Figure 3.S2c**), implying some stereotypical decision-related dynamics (whereas that is less constrained in factor analysis; **Figure 3.S4d**). Consistent with the strong epoch-dependent change in dynamics, we find that epoch-dependent matrices **A** and **C** were required to obtain good fits (**Figure 3.1e; Figure 3.S2b**).

3.2.2 Decoders operating on the shared activity space predict behavioral variability

We next tested whether the representation in the shared activity space would yield more accurate predictions of behavioral variability. We first consider decoding trial


Trial type decoding
a
b
Tiral counts
20
0
-2
0
2
LDA proj. (a.u.)
Tiral counts
25
0
-3
0
2
TLDS-LDA proj. (a.u.)
Accuracy TLDS-LDA
1
0.5
0.5
1
Accuracy LDA
Reaction time prediction
c
d
First lick time (ms)
400
0
-2
0
2
LDA proj. (a.u.)
First lick time (ms)
400
0
-3
0
2
TLDS-LDA proj. (a.u.)
Contra. corr. TLDS-LDA vs RT
0.5
0
-0.5
-0.5
0
0.5
Ipsi. corr. TLDS-LDA vs RT
0.56
explained variacne
0.10

types. Decoders trained either on the FNS or the SAS could separate single trials from different trial types (**Figure 3.2a**; FNS, top; SAS, bottom) but decoders running on SAS were more accurate ($p < .001$, two-sample t-test, **Figure 3.2b**). We then considered more subtle predictions, and chose to attempt to explain the behavioral variability in time of movement onset. The projection of the activity on FNS trial-type decoders did not explain much behavioral variability (**Table 3.S5**, reaction time fits for FNS LDA score vs SAS LDA score; significant correlation sessions, $p < .05$, FNS, 5 sessions; SAS, 7 sessions). However, the projection of decoders operating on SAS were significantly correlated with movement onset (**Figure 3.2c**; FNS, top; SAS, bottom). Interestingly, there was a significant correlation for ipsi trials (r_s = -.29; $p < .05$), but not for contra trials (r_s = -.01; $p > .05$). We found that this was the case for most sessions, where the TLDS-LDA score correlated with behavioral variability

Figure 3.2 *(preceding page)*: Trial identity decoded from TLDS model correlates with trial-by-trial variability in behavioral reaction time and performance on single trials.
a. LDA score in shared-activity space shows clear separation of trial types but overlapped in neural space (contra. trial, blue; ipsi. trial, red; 15-unit simultaneously recording session). **b.** Across sessions, the neural-mode LDA score (based on the dynamics in a 67-ms time bin before onset of response) outperforms others (based on the dynamics in a 350-ms time bin before onset of response) in full-neural space when decoding trial types (each circle represents a simultaneously recorded session; color of each circle represents the amount of explained variance captured by TLDS model; error bar, sem. across trials). **c.** LDA score in shared-activity space shows strong correlation with reaction time to first lick in single trials during a single session (contra. trial, blue; ipsi. trial, red; 15-unit simultaneously recording session). Correlation was asymmetric and stronger for ipsi. trials. **d.** Across session, neural-mode LDA score can predict reaction times to first lick in single trials for a specific trial type (significant correlation, filled circle, $p < .05$; otherwise, empty circle; color follows same schematics as in **b**).

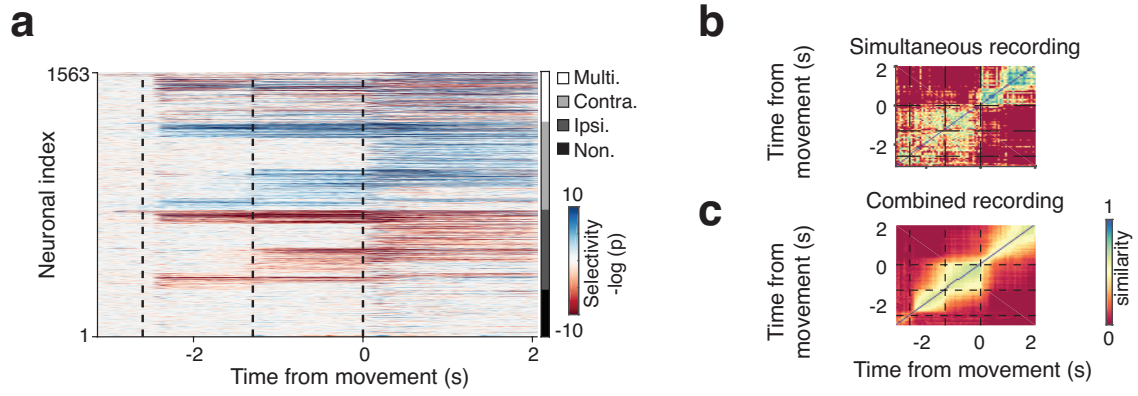


Figure 3.3: Neural activity of anterolateral motor cortex neurons exhibits a discontinuity of selectivity at response.
a. Color plot of instantaneous trial-type selectivity, displayed as a negative log of p-value, (blue: contra-preferring; red: ipsi-preferring). A notable fraction of neurons exhibited switches of trial type selectivity in time (white, multiphasic selective neuron), implying the discontinuous dynamics of selectivity at a single-unit level. The rest showed consistent polarity of trial type selectivity (black, non-selective; dark gray, ipsi-monophasic selective; light gray, contra-monophasic selective). **b-c**. Similarity of coding direction (linear discriminant analysis of trial type, LDA) across time. LDA is similar within epochs, but different from sample-delay epoch to response, indicating the discontinuous dynamics of selectivity at the population level. **b.** Analysis using simultaneous recorded units (n = 15) and trials; **c.** analysis using all non-simultaneous recorded units (n = 1,563).

in only one of the trial types, either contra. or ipsi. (**Figure 3.2d**).

3.2.3 Strong switches of neural dynamics at response in ALM

Many ALM neurons exhibited complex dynamics in time (**Figure 3.3a**). For example, 23% of ALM neurons (n = 367) showed a switch in their trial-type selectivity from the sample-delay epochs to the response epoch. We now consider these

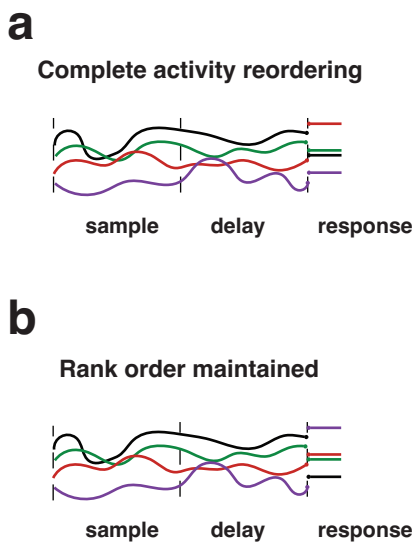


Figure 3.4: Two hypothesized models of trial identity dynamics.
Two hypothesized models for the relation between the dynamics of individual trials across a switch in population dynamics. Schematic representation of population activity over time, different colors indicate different individual trials. Due to the strong dissimilarity of trial type decoders across behavioral epochs, one could assume that there would be a complete loss of trial identity across a transition in the dynamics of the network (**a**), resulting in no relation between the ranks of activity levels before and after the transition. Alternatively, these dynamics could switch, but information prior to the switch will be maintained or completely flipped in order (where the absolute value of rank-order correlation remains high) in the deviation of a single trial from the average activity (**b**).

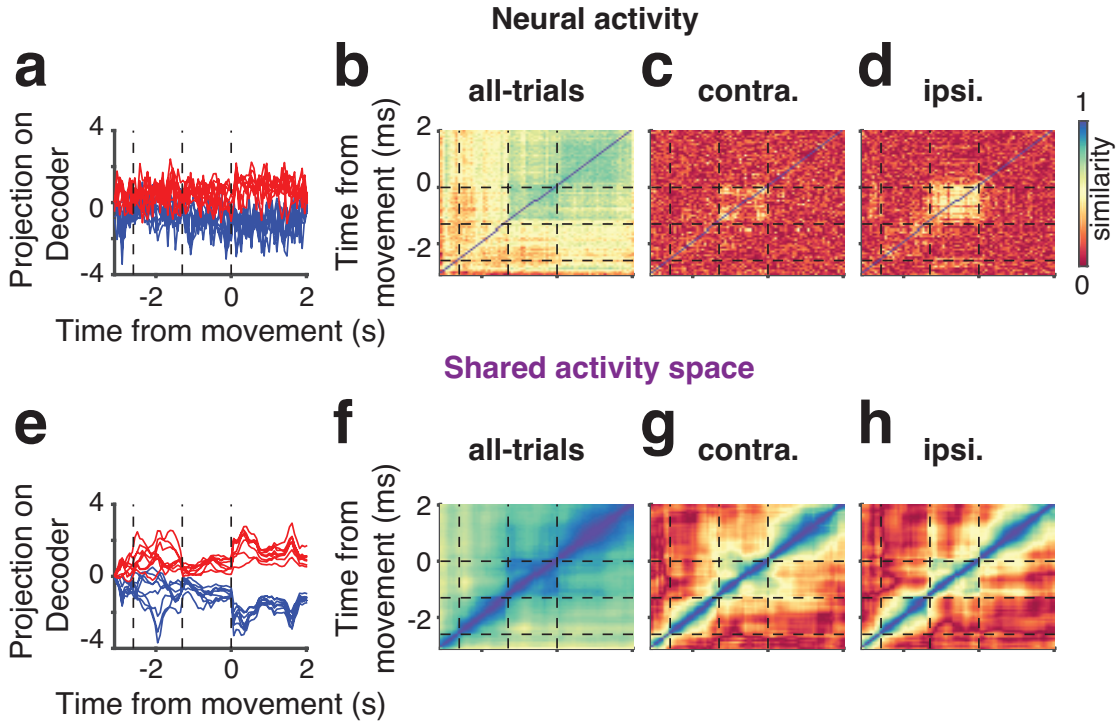


Figure 3.5: Maintenance of trial identity across dynamical transition revealed by time-varying linear dynamical systems model.
a-d. Dynamics of trial identity are discontinued across behavioral epochs in neural space. **a**. The dynamics of an instantaneous neural LDA score. The projection of activity in full-neural space onto the 1^{st} discriminant component exhibits a robust separation of trial type maintained from epoch to epoch, but rank correlation is weak within the same trial type. **b**. Rank correlation of the instantaneous LDA score is strong across epochs when trial types are pooled together. **h-i**. Rank correlation of the instantaneous LDA score is weak even within a behavioral epoch when trial types are considered separately (contra-trial, **c**; ipsi-trial, **d**), indicating that trial identity is not maintained in a rank-like order across the dynamics transition. **e-h**. Dynamics of trial identity are continuous across behavioral epochs in shared-activity space. **e-h**. the same as **a-d** in shared-activity space. When calculated directly in the shared-input space, the instantaneous neural-mode LDA score is of high similarity even across behavioral epochs in the same trial type. Rank correlation of the instantaneous LDA score is strong within a time window of hundreds of milliseconds across behavioral epochs, even for data taken separately for each trial type (contra-trial, **g**; ipsi-trial, **h**), indicating that trial identity is maintained by population activity, and can be uncovered by the TLDS model.

switches at the population level and how they affect the latent dynamics. Examining the relations between linear discriminant analysis (LDA) decoders of trial-type at separate times, we find that decoding directions were similar among different times within one behavioral epoch, and had a strong switch in-between the sample-delay and response epochs (**Figure 3.3b**, 6 - 18 units per session, 19 sessions; **Table 3.S1**). The same held true for pooled across-session data (**Figure 3.3c**) and for shuffled data (**Figure 3.S1d**) (Li et al., 2016). Notably, LDAs in sample-delay epochs were almost orthogonal to those calculated very soon after the response, which could result from two possibilities: (1) many ALM neurons (n = 367) showed multiphasic-selectivity and switched their preference significantly at response times or that (2) mutually exclusive subpopulations of neurons encoded the trial type at different behavioral epochs (the dynamics would be discontinued at response). Surprisingly, we found that most variance (84%) for switching neural dynamics occurred in the subpopulation of monophasic-selective neurons (n = 929; **Figure 3.S1e**), which supports more strongly the second explanation. This raises the question of whether any neural dynamics could be continuous in time across such a dramatic switch in network dynamics, and accordingly, how much information is maintained from the decision epoch to post-decision times.

3.2.4 Continuous trial identity signals in the shared activity space

Consider two extreme options for the behavior of dynamics across a strong dynamical switch. In the first option, there is a complete erasure of information across the switch. In this case, when considering the dynamics in different states and trials before their switch, characterized such as their rank among all the trials, will yield no information regarding the state after the switch (**Figure 3.4a**). Alternatively, the dynamics could have a strong, but orderly switch which would preserve much information. In this scenario, the rank of a particular trial among others may well be preserved despite the strong switch in dynamics, for instance (**Figure 3.4b**). These two hypotheses are of course extremes. Here, we quantify this general tendency by examining the rank ordering of trials before and after the switch between the delay and response periods. We perform this analysis both for the full and for the shared neural space.

We first tested the hypotheses based on the projection of simultaneous neural dynamics to the instantaneous LDA (neural LDA score) (**Figure 3.5a**). The neural LDA score was noisy in time but exhibited a separation across trial type from sample to response epochs, which kept the rank of trial type consistent in time (**Figure 3.5b**). Nevertheless, we found the rank correlation of neural dynamics was weak across epochs within the same trial type (**Figures 3.5cd**). This could result from

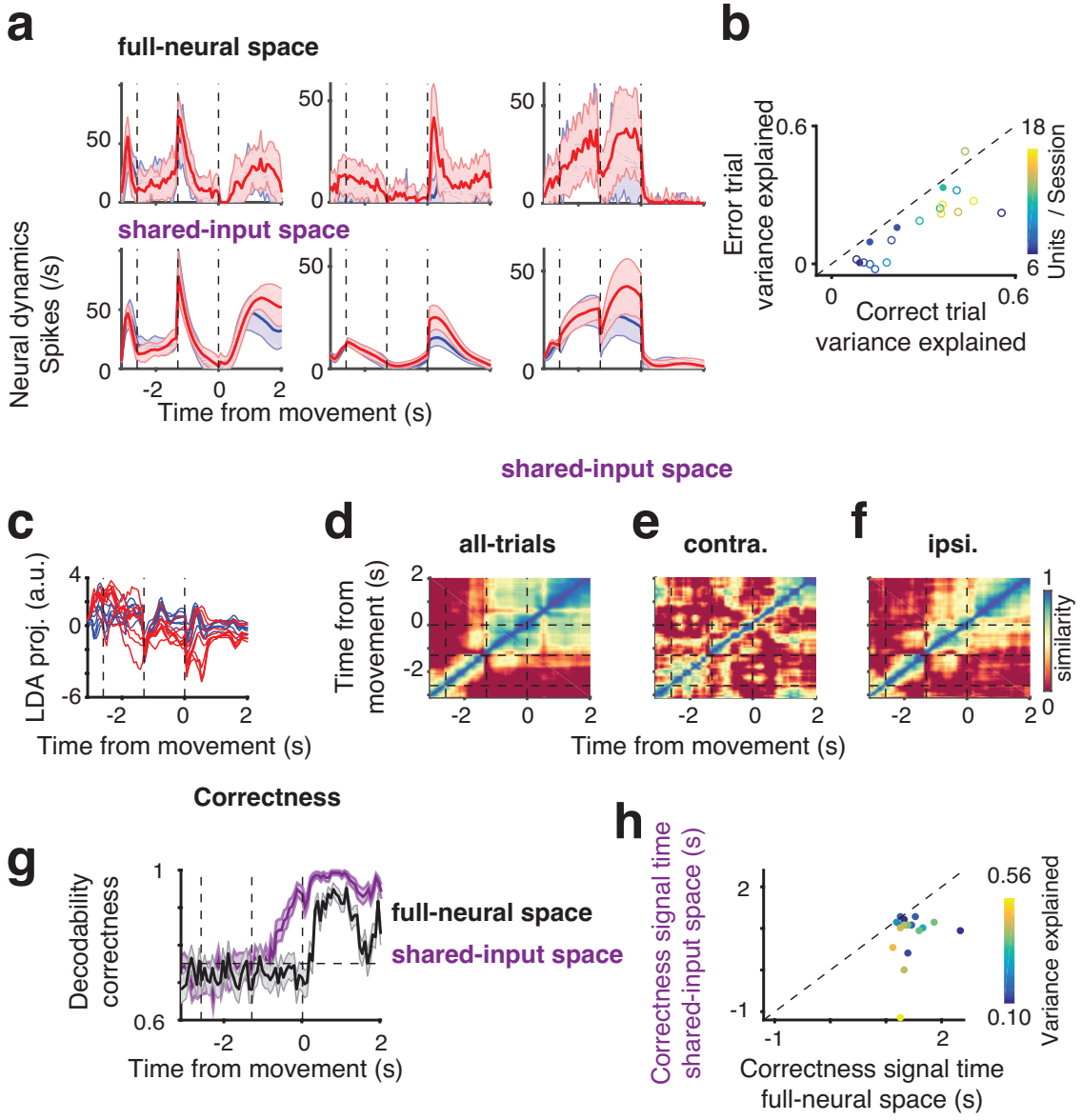
the level of noise in FNS. We therefore examined rank correlation in FNS temporally smoothed by different boxcar time windows (**Figures 3.S3ab**), which slightly increased the rank correlation of neural LDA scores across epochs (**Figure 3.S3c**).

The failure to uncover continuity of rank dynamics could result also from the strong independent inputs onto single neurons. We thus assumed that its continuous presentation could be maintained in the dynamics of shared activities. We found that the projection of neural-mode dynamics in SAS to the instantaneous LDA (neural-mode LDA score) was clearly separated across trials (**Figure 3.5e**) and strongly correlated across time (**Figure 3.5f**), even within the same trial type (**Figures 3.5gh**); a comparison with neural LDA score in boxcar-smoothed FNS, **Figure 3.S3c**).

In summary, we found that the rank dynamics of a single trial were inconsistent and chaotic in FNS, due to the independent drives to each neuron, but highly ordered in SAS, which supports the second hypothesis and indicates distinguishable neural dynamics, as a trial identity signal, for each trial.

3.2.5 Error trials

The source of errors in decision making trials is often unclear, since the final action depends on the robust integration of all sources of the neural signals (eigenvalues were all close to one, **Table 3.S6**), and many factors such as noise during the process of perception, or choice generation, or memory could have an impact on decisions. Therefore, it is an emerging interest in systems neuroscience to track down the source


a
full-neural space
shared-input space
Neural dynamics
Spikes (/s)
50
0
-2
0
2
Time from movement (s)
b
Error trial
variance explained
Correct trial
variance explained
0.6
0
18
6
Units / Session
shared-input space
c
LDA proj. (a.u.)
4
0
-6
Time from movement (s)
d
all-trials
Time from
movement (s)
e
contra.
f
ipsi.
1
similarity
Time from movement (s)
Correctness
g
Decodability
correctness
1
0.6
full-neural space
shared-input space
Time from movement (s)
h
Correctness signal time
shared-input space (s)
Correctness signal time
full-neural space (s)
2
-1
0.56
0.10
Variance explained

of error signals, as the technologies to observe and decode the single-trial dynamics of the relevant circuits has become available (Kiani et al., 2015). Here, we examined the structure of the dynamics in correct and in error trials. Surprisingly, we found that the SAS could still predict substantial variability in a leave-one-neuron-out test even during error trials (**Figure 3.6a**). Although the fraction of explained variance was smaller in error trials ($p < .001$; **Figure 3.6b**), its value correlated strongly with the explained variance in control trials ($r_s = .84$, $p < .001$), indicating that similar neural dynamics underlie the formation of both correct and error trials, but with different levels of noise.

Figure 3.6 *(preceding page)*: TLDS model still has predictive power in error trials and reveals failure of trial identity maintenance in error trials. **a.** Neuronal activity in error trials can be explained by other neurons in the same recording session using TLDS model ***estimated from correct trial data only*** (15-unit simultaneously recorded session). Top: observed neural activity in error trials; bottom: prediction of neural activity using the leave-one-neuron-out methods based on TLDS model fit given correct trials. Shaded area, *std.* across trials. **b.** Across sessions, a substantial fraction of variability in error trials can be explained by using an identical TLDS model as for correct trials, implying that an error trial would be reproduced from a similar network as a correct trial, while the noise level is comparatively large (each circle represents a simultaneously recorded session; color of each circle represents the number of simultaneously recorded neurons; significant difference, empty circle, $p < .05$; otherwise, filled circle). **c-f.** Same analyses of LDA scores in shared-activity space using error trials as those in **Figure 3g-j**. Of note, rank-order correlation of the instantaneous LDA score is weak across behavioral epoch even for a trial type (contra-trial, **e**; ipsi-trial, **f**), indicating that trial identity would be a complete loss even in error trials. **g.** The neural decodability of a trial-correctness signal in full-neural space (black line; shaded area, sem. across trials) was delayed compared to the signal in shared-activity space (purple line). **h**. This holds true across sessions (each circle represents a simultaneously recorded session; color of each circle represents the amount of explained variance captured by a TLDS model in correct trials); time is aligned to the onset of response.

If all trials follow the seemingly identical dynamics, what drives error trials to make an error? We examined this through the dynamics of trial identity. In error trials, the trial identity could be disrupted randomly in time. We found that the overall correlation was weak and discontinuous in error trials across epochs (**Figures 3.6c-f**).

Importantly, we found that an apparent misclassification of trial identity could happen early at the onset of a delay epoch (**Figure 3.6d**), which suggests that a choice formed in the sample epoch was perhaps not correctly stored during the delay. We therefore hypothesized that there could be some internal representation of the correctness of a trial early on. We computed the instantaneous neural decodability of trial-correctness in both FNS and SAS (**Materials and Methods**) and found that the dynamics in SAS exhibited early separation of correct versus error trials, showing neural prediction of trial-correctness in sample-delay epochs, while in FNS, this separation was delayed until a response was made (**Figure 3.6g**). We note that this is a prediction of trial-correctness without access to what the original identity of the trial was, therefore it is not sufficient to decode trial type. Across sessions, the dynamics of neural modes showed an early prediction of the trial-correctness ($p < .001$; **Figure 3.6h**).

3.3 Discussion

Identification of a decision variable (DV) is the key to neural mechanism of decision making (Gold and Shadlen, 2007). This was demonstrated by population codes of the simultaneous neural activities (e.g. neural LDA score) (Beck et al., 2008). In line with other studies, we found that the dynamics of DV in full-neural space (FNS) was discontinuous at response time (Li et al., 2016). Nevertheless, the recent studies showed that pre-sample neural activity could contain the information from previous decisions, indicating that DV had to be sustained even after the response (Morcos and Harvey, 2016). Therefore, we studied the neural dynamics in a shared-activity space (SAS) and discovered the continuous signal of DV for decision and post-decision neural dynamics and behaviors. Furthermore, our definition of DV in SAS (neural-mode LDA score) carried the robust information of trial identity that could differentiate among variabilities of the single trials in great details (e.g. reaction time, performance and correctness). Importantly, our analysis of the neural activities in SAS demonstrated quantitatively that (1) the dynamics of DV followed the nearly non-leaky drift-diffusion model (Ratcliff and Smith, 2004; Ratcliff and Starns, 2009; Kiani et al., 2008) (eigenvalues of latent dynamical systems are close to ones; **Table 3.S6**), (2) the dynamics of DV in errors could follow the same mechanism of that in correct trials, except the amount of noise was larger, and (3) the neural signal could also provide the early prediction of choice correctness, which mimicked a confidence signal in decision making (Wei and Wang, 2015). The DV and confidence signal were

both in SAS, although in different directions, they were partially correlated; notably, the correlation was weak in FNS (data not shown). To our knowledge, this is the first work to quantitatively examine the prediction power of error-trial neural dynamics based on a model estimated by the correct trials.

Neural dynamics in SAS found the robust signals of trial identity that (1) were shared across multiple neurons, and (2) demonstrated both the neural and behavioral variabilities (e.g. reaction time, performance and correctness) on single trials. The temporal robustness of the neural signal could be explained by (1) shared inputs distributed across many neurons in a local circuit and (2) multiple copies of the similar neural circuits on two hemispheres (Li et al., 2016). Since the neural signal in SAS was robustly maintained for several seconds (>3 s on average; **Table 3.S6**) in post-decision, such a neural signal could show up in pre-sample activity in the following trial, which explains how a previous trial type is identified in pre-sample neural activity.

Importantly, these neural dynamics were not simply equivalent to the de-noised input onto single neurons. This was supported by the quantitative analysis in our study and by the theoretical models (Druckmann and Chklovskii, 2012). The independent non-informative input in FNS (even without noise) could drive the seemingly chaotic behaviors of single neurons (usually assumed by random networks), while these dynamics would reside mostly in the null-space orthogonal to the coding space (Druckmann and Chklovskii, 2012; Li et al., 2016). Nevertheless, modeling neural dynamics

in SAS was difficult if the single neuronal activity was strongly locked to a behavioral epoch in time (**Chapter 2**). One could determine such a transition of the neural dynamics using the Hidden Markov Model (Petreska et al., 2011), or alternatively, as we did, by using a time-varying version of the latent dynamical systems.

Multi-electrode and optical imaging recordings provide simultaneous monitoring of activity from tens to hundreds of neurons, and thus enable us to probe the statistical structure of neural population activity. State-of-the-art statistical tools are required to help us explore the informative neural dynamics in high dimensional space. Our analysis, using the latent dynamical systems and leave-one-neuron-out estimations, not only revealed the applicability of latent-space analysis in neural data to uncover the task-relevant neural space (besides a simple dimensionality reduction), but also provided the insight into the discovery of the neural signal to the continuous presentation of DV and confidence signal that underlay the behavioral variabilities. Our current study showed that the reaction time were predictable from DV, which, however, only applied to an exclusive trial type. In future research, one could examine whether this holds true for simultaneous recordings with a larger number of the units using calcium imaging, for which, one should consider the independent slow dynamics of each unit in TLDS fit.

3.4 Materials and Methods

3.4.1 Electrophysiological recordings

Mice were trained to perform a delayed version of a tactile discrimination task (1.3-s sample and 1.3-s delay). Mice reported the position of a pole (anterior or posterior) by directional licking (lick-left, trial type A; or lick-right, trial type B) after a delay period. Electrophysiological recordings were performed on the left-hemisphere anterolateral motor cortex (ALM) using 32-channel NeuroNexus silicon probes or 64-channel Janelia silicon probes. The details of electrophysiology and spike sorting were described in (Li et al., 2015) and (Guo et al., submitted).

We excluded trials with early licking, and for non-simultaneous recording data, we selected neurons with >20 trials for each type (contra. and ipsi.) (**Figures 3.3a-c**). Cells were previously classified as the fast-spike interneurons (n = 243) and pyramidal cells (n = 1,320) in (Li et al., 2015) and (Guo et al., submitted). Neuronal depths were at 48 - 1,329 um (32-channel) and 178 - 1,052 um (64-channel).

The simultaneous recording sessions were collected as those (1) that have > 5 units and (2) where the number of correct trials is more than the double the number of units (8 sessions recorded using 32-channel; 11 sessions using 64-channel). The fraction of pyramidal cells (see classification of cell types in (Li et al., 2015) and (Guo et al., submitted)) in the simultaneous recording sessions ranged from 64% to 100%, with an average of 79%; the depth was from 219 to 1,044 um; the difference of that in

a single session was from 256 to 645 um, with a mean at 478 um. A detailed summary is shown in **Table 3.S3**.

An example session (Session #11; **Table 3.S3**) shown in the text was a 15-unit simultaneous recording at depths of 297 - 863 um and was one of the two sessions without interneurons.

3.4.2 Single neuron analysis

For all analyses, we binned neural activity using a 67-ms discrete time window; except that in **Figures 3.S3ab**, where we computed spike counts in a 150-ms (**Figure 3.S3a**) and 250-ms (**Figure 3.S3b**) in 10-ms steps. To compare the single trial variability in different neural spaces, we computed std. (the standard deviation) of neural dynamics within the same trial type; but **Figure 3.S1** shows sem. (the standard error of the mean).

To measure the dynamics of selectivity, we performed two-sample t tests with neural activity over 67 ms discrete bins (**Figure 3.S1b**). We defined a neuron as monophasic if it had consistent polarity of selectivity ($p < .05$) for >335 ms (5 continuous bins); a neuron was multiphasic if it had a switch of selectivity ($p < .05$) with the periods of selectivity being at least 335 ms. The rest of the neurons were considered to be nonselective neurons. We classified monophasic-selective neurons into contra.- and ipsi.-preferring cells, according to the trial type for which they had higher activity ($p < .05$, two-sample t-test).

3.4.3 Coding directions and neural dynamics in coding directions

To determine the coding directions by trial type, we applied linear discriminant analysis (LDA) to neural dynamics grouped into 67-ms non-overlapped bins. The optimal LDA decoder, $\mathbf{l}_t$, was computed separately for each time bin, t, using trials in correct responses, where we indexed contra. trials as zeros and ipsi. trials as ones. The neural dynamics projection to coding directions (neural LDA score) was therefore computed as $s_t = \mathbf{l}_t^T \mathbf{r}_t$, where $\mathbf{r}_t$ is the vector of neural activity in time bin t. This was based on our trial type index, $s_t < 0$ in contra. and $s_t > 0$ in ipsi. trials. The same principle was applied to estimating LDA trial-type direction and its score in other neural spaces. The correlation of LDA scores across time, t, and, t', was performed using Spearmen's rank correlation, $r_s(t,\ t')$ (which was not significantly less than zero; $p < .05$), and across epochs, s, and, s', which was $< r_s(t,\ t') >_{t \in s; t' \in s'}$.

Performance in the full-neural space was computed based on the neural activity in a 350-ms time bin before the onset of response using 10-fold cross-validation (**Figure 3.2b**; error bar, std.); and in the shared-activity space, performance was computed based on the instantaneous neural dynamics of a 67-ms time bin before the onset of response.

To achieve robust estimation given the large number of neurons, we applied a sparse version of LDA (Guo et al., 2007), with normalization of $||\mathbf{l}_t||_2 = 1$ (where

$||\cdot||_2$, the L^2 norm of the vector). In this case, the LDA coefficient would be close to zero if a neuron fired at a low rate or contributed little to coding trial type. We measured the similarity among coding directions across time, t, and, t', as $\mathbf{l}_t^T\mathbf{l}_{t'}$. We did not observe any $\mathbf{l}_t^T\mathbf{l}_{t'} < 0$.

We also performed similar LDA on neural dynamics to determine the coding directions of trial-correctness (correct versus error trials) and similar computation of LDA scores in trial-correctness directions within different neural spaces. The performance (decodability of trial-correctness) was computed based on instantaneous neural dynamics in discrete 67-ms time bins, using 10-fold cross-validation (shaded area, sem.; **Figure 3.6g**). The chance level of the reward was computed as a fraction of correct trials (**Table 3.S5**). The trial-correctness signal times were computed as those when performance was significantly above the chance ($p < .05$, t test; **Figure 3.6h**).

3.4.4 Time-varying linear dynamics systems model and neural mode dynamics

A time-varying linear dynamics systems (TLDS) model is an extended model of linear dynamics systems (Roweis and Ghahramani, 1999; Ghahramani and Hinton, 1996a; Buesing et al., 2012; Gao et al., 2015; Kao et al., 2015; Macke et al., 2011), with explicit modeling of the dynamics of the latent connectivity, $\mathbf{A}$, and the projection, $\mathbf{C}$. The goal of TLDS was to reproduce the highly-correlated neural activity ($\mathbf{r}(t) \in \mathbf{R}^N$)

in the most possible low-dimensional neural-mode space ($\mathbf{x}(t) \in \mathbf{R}^M$, $M < N$), where neural modes followed explicitly the linear dynamics (**Equations 3.1 and 3.2**), which were equivalent to maximize the joint probability

$$P(\mathbf{x}(t), \mathbf{y}(t)) = P(\mathbf{x}(1)) \prod_{t=2}^{T} P(\mathbf{x}(t)|\mathbf{x}(t-1)) \prod_{t=1}^{T} P(\mathbf{r}(t) - \mathbf{r}_0|\mathbf{x}(t)),$$

where $\mathbf{r}(t) - \mathbf{r}_0|\mathbf{x}(t) \equiv \bar{\mathbf{r}}(t) \sim N(\mathbf{C}(s)\mathbf{x}(t), \mathbf{Q}_{ext}(s))$; $\mathbf{x}(t)|\mathbf{x}(t-1) \sim N(\mathbf{A}(s)\mathbf{x}(t-1), \mathbf{Q}_{int}(s))$, and the initial state of the neural mode $\mathbf{x}(1) \sim N(\mathbf{x}_0,\ \mathbf{Q}_0)$.

The parameter set Θ included the amplitudes of external independent inputs $\mathbf{Q}_{ext}(s) = \sigma^2_{ext}(n)$ onto each neuron ($1 \leq n \leq N$; N, the number of neurons in simultaneous recordings), the amplitude internal inputs $\mathbf{Q}_{int}(s) = \sigma^2_{int}(m)$ in each neural mode ($1 \leq m \leq M$; $M < N$, the latent dimension), the latent connectivity, $\mathbf{A}(s) \in \mathbf{R}^{M \times M}$, and the projection, $\mathbf{C}(s) \in \mathbf{R}^{N \times M}$ for each epoch (i.e. pre-sample, sample, delay, and response), and those associated with the initial state of the neural mode $\mathbf{x}_0$, $\mathbf{Q}_0$ in each trial.

The estimation of TLDS followed the similar Expectation-Maximization step previously given by Ghahramani and Hinton (1996a). The expectation step was identical and the maximization step was adopted for TLDS as:

$$\mathbf{C}^*(s) = \Big(\sum_{t=t(s-1)}^{t(s)} \bar{\mathbf{r}}_t \mathbf{x}_t^T \Big) \Big(\sum_{t=t(s-1)}^{t(s)} \mathbf{x}_t \mathbf{x}_t^T | \{\bar{\mathbf{r}}_t\}_{1 \leq t \leq T} \Big)^{-1}, \tag{3.3}$$

$$\mathbf{Q}^*_{ext}(s) = \frac{1}{t(s)+1-t(s-1)} \sum_{t=t(s-1)}^{t(s)} (\bar{\mathbf{r}}_t - \mathbf{C}^*(s)\mathbf{x}_t)\bar{\mathbf{r}}_t^T, \tag{3.4}$$

$$\mathbf{A}^*(s) = (\sum_{t=t(s-1)+1}^{t(s)} \mathbf{x}_t\mathbf{x}_{t-1}^T|\{\bar{\mathbf{r}}_t\}_{1\le t\le T})(\sum_{t=t(s-1)+1}^{t(s)} \mathbf{x}_{t-1}\mathbf{x}_{t-1}^T|\{\bar{\mathbf{r}}_t\}_{1\le t\le T})^{-1}, \tag{3.5}$$

$$\begin{aligned}\mathbf{Q}^*_{int}(s) = \frac{1}{t(s)-t(s-1)}(&\sum_{t=t(s-1)+1}^{t(s)} \mathbf{x}_t\mathbf{x}_t^T|\{\bar{\mathbf{r}}_t\}_{1\le t\le T} \\ &-\mathbf{A}^*(s)\sum_{t=t(s-1)+1}^{t(s)} \mathbf{x}_{t-1}\mathbf{x}_{t-1}^T|\{\bar{\mathbf{r}}_t\}_{1\le t\le T}).\end{aligned} \tag{3.6}$$

where $t(s)$ represents the final time point of epoch s and $t(0) = 1$. Here, we constrained the estimation of input matrices $\mathbf{Q}_{ext}(s)$ and $\mathbf{Q}_{int}(s)$ to be diagonal after each iteration of maximization. Further, we computed the dynamics of neural mode in SAS as $\mathbf{x}_t|\{\bar{\mathbf{r}}_t\}_{1\le t\le T}$ based on the expectation step.

Since only a small fraction of trials were errors (**Table 3.S2**), we cannot perform the TLDS fit directly on error trials. Alternatively, the estimation of neural dynamics in error trials was based on the TLDS fit from correct trials (**Figure 3.6**).

3.4.5 Leave-one-neuron-out estimation and optimal dimension of the shared-input space

To determine the performance of the TLDS fit, we computed the variance explained below using an adapted leave-one-out procedure in cross-validation. The dynamics of each neural mode in the SAS modeled shared dynamics of many neurons,

and little is influenced by kicking one neuron out of the estimation. Here we assumed that $\mathbf{x}_t|\{\bar{\mathbf{r}}_t\}_{1\leq t\leq T} \cong \mathbf{x}_t|\{\bar{\mathbf{r}}_t\}_{1\leq t\leq T}$ to be the leave-one-neuron-out (LONO) estimation (the ith neuron was kicked out) of the neural-mode dynamics. One would then compute the estimation of the dynamics of ith neuron as $\hat{r}_t^i = \mathbf{C}^i(s)\mathbf{x}_t|\{\bar{\mathbf{r}}_t\}_{1\leq t\leq T} + r_0^i$, where $\mathbf{C}^i$ is the ith row of the projection matrix $\mathbf{C}$. The explained variance,

$$R^2 = 1 - \frac{< ||\mathbf{r}_t - \hat{\mathbf{r}}_t||_2^2 >_t}{< ||\mathbf{r}_t - \mathbf{r}_0||_2^2 >_t},$$

thus measured the goodness-of-fit (where $||\cdot||_2$, the L^2 norm of the vector; $< \cdot >_t$, averaged over time and trials), and determined to which degree the dynamics of a single neuron can be represented by other neurons in the same population under the assumption of a particular TLDS model. We called this procedure the LONO estimation of TLDS fit.

To determine the optimal dimension of SAS, M, we performed a LONO estimation of TLDS fit using 10-fold cross-validation for each possible dimension, $1 \leq M \leq N-2$ (**Figure 3.S2b**). The amount of the variance explained first increased (underfitting region) then saturated or even decreased (overfitting region) as a function of the SAS dimension. To avoid overfitting (where a neural mode was essentially projected back to a single neuron), we picked the smallest dimension that reached >90% of the maximum amount of variance explained. Moreover, to establish necessity for epoch-dependent dynamics of matrices $\mathbf{A}$ and $\mathbf{C}$, we analyzed the amount of variance

explained in TLDS fit where **A** and **C** were constant matrices (constant **A-C** fit, blue; epoch-dependent **A-C** fit, purple; **Figure 3.S2b**).

3.4.6 Reaction time correlations

We correlated the neural dynamics in coding directions with reaction times of the first lick using Spearmen's rank correlation. The neural dynamics in FNS were estimated as the neural LDA score in a 350-ms time window before the onset of response (**Figure 3.S5e**), and in SAS these were estimated as the instantaneous neural mode LDA score in a 67-ms time window before the onset of response (**Figure 3.2d**). We used bootstrap (1,000 replications) to examine the significance of the correlation ($p < .05$).

3.5 Supplementary

3.5.1 Supplementary figures

Figure S1. (Supporting figure for Figure 3.3) Neural activity in ALM exhibits a switch of selectivity at response in the 1^{st} PCA space

a. Exemplary single neural dynamics recorded from the anterolateral motor cortex (ALM) exhibit diverse dynamics and trial type selectivity. Left: a monophasic-selective neuron (contra-preferring). Color indicates trial type: contra. (blue) or ipsi. (red), top: raster plots, bottom: trial averaged activity over trial types (shaded area, sem. across trials). Middle: a multiphasic-selective neuron, i.e., a neuron that exhibits a switch of trial type preference. Right: a non-selective neuron. **b.** Fraction of variance and content in principal component directions. The 1^{st} principal component represents mostly the dynamics of temporal components across neurons. c. Neural dynamics projected to the 1^{st} LDA direction of trial type, where the LDA decoder was computed based on averaged neural activity in sample (left), delay (middle) or response (right) epochs. **d.** Same as **Figure 3.3b**, except for a shuffled trial. **e.** Projection onto the 1^{st} principal component over time, separately for each trial type (lick-left red, lick-right blue). Data separated into populations according to neuronal selectivity. From left to right: contra-preferring monophasic-selective, ipsi-preferring

monophasic-selective, multiphasic-selective, and non-selective neuronal populations. Interestingly, a switch of selectivity is present not just for the multiphasic-selective population, but also for contra-preferring monophasic-selective population.

Figure S2. (Supporting figure for Figure 3.1) Fitting details of time-varying linear dynamical system models

To determine the importance of a time-varying hypothesis of linear dynamical system (LDS) fit, we refitted neural dynamics using the LDS model with constant matrices **A** and **C**. **a.** The time-independent LDS can reproduce neural activity in simultaneous recordings using the other neural activities from the same pool. Top: Observed neuronal activity from 15-unit simultaneous recording session, bottom: prediction of neuronal activity on prediction data where the activity of that neuron is unobserved. However, the explained variance is low in a time-independent version of the LDS model. **b.** Comparison of LDS fit in time-varying (purple line and circles; error bar, std estimated from 10-fold cross-validation) and time-independent conditions (blue line and circles) across 19 simultaneous recording sessions. The explained variance increases with the dimension of shared input space and is consistently high in TLDS. **c.** We determined the optimal dimension of TLDS fit to be the minimum dimension that explained $> .90$ peak variance in all fits. The optimal dimension is concentrated at 3 or 4 across sessions, regardless of the dimension of full-neural space.

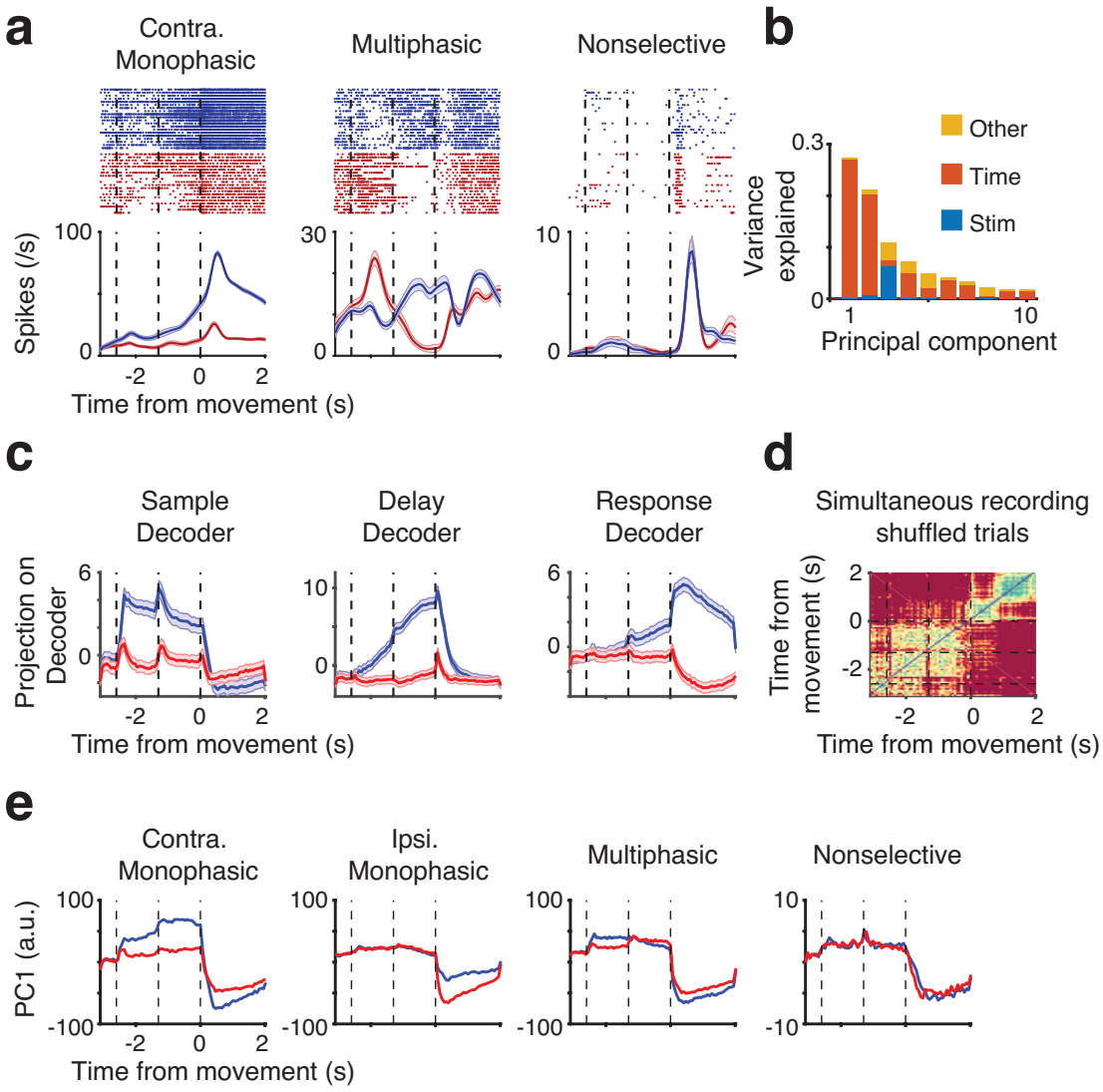


Figure 3.S1: Neural activity in ALM exhibits a switch of selectivity at response in the 1^{st} PCA space.

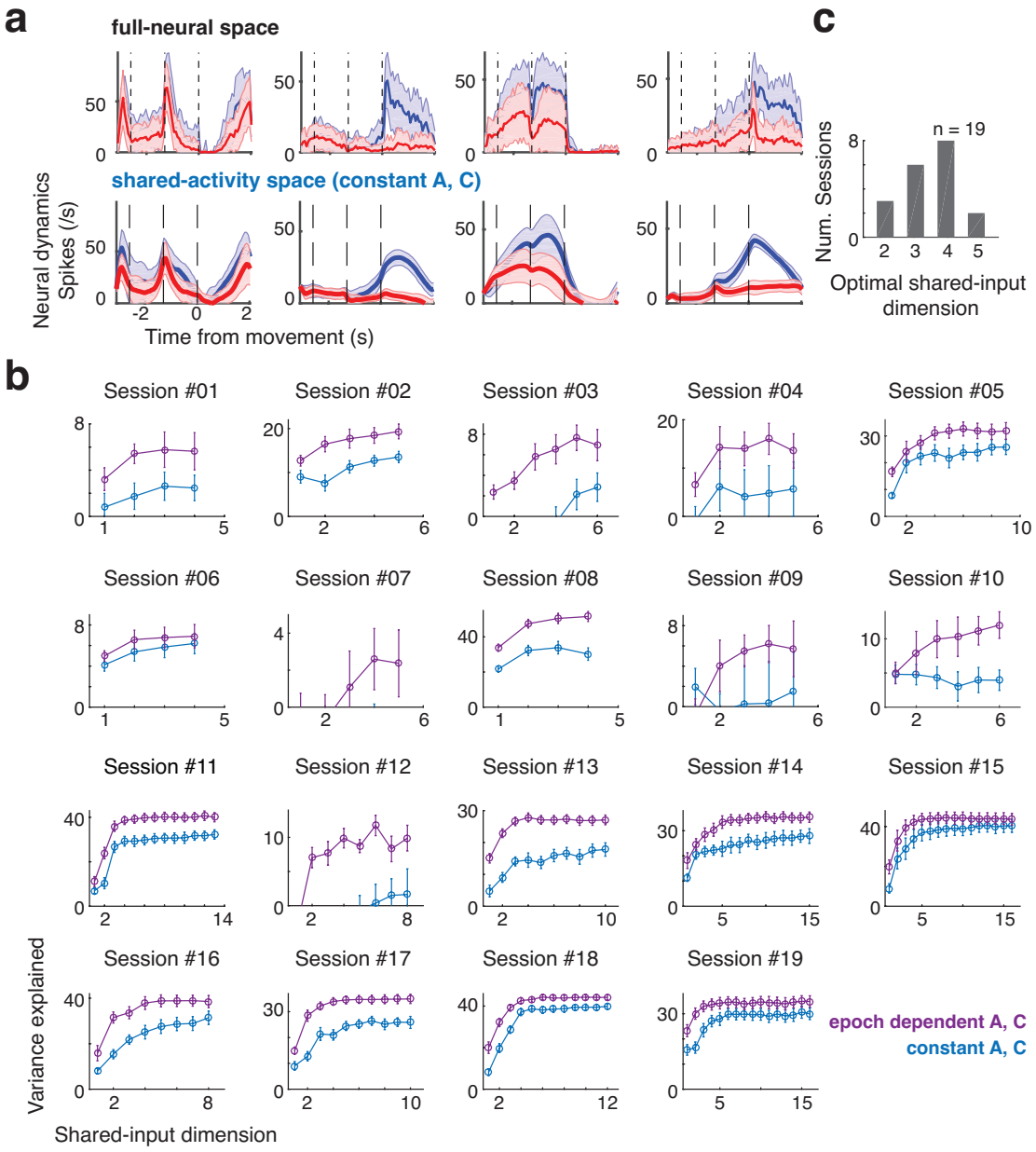


Figure 3.S2: Fitting details of time-varying linear dynamical system models.

Figure S3. (Supporting figure for Figure 3.5) Rank correlation of trials is higher in shared-input space than that in boxcar-smoothed full-neural space

a. Same analysis as **Figure 3.5a-d** using temporally boxcar-smoothed neural activity in full-neural space (time window is 150 ms). **b.** The same analysis as **a** with boxcar time window at 250 ms. The rank correlation within local time window increases compared to discrete binned neural activity (**Figures 3.5a-d**); however, the rank correlation across epochs does not improve notably. **c.** Comparison of rank correlation (similarity index) in smoothed full-neural space (250-ms boxcar) and that in shared-input space (each circle represents a simultaneously recorded session; error bar, std .of rank correlation using bootstrap; color, average rank correlation across epochs; black, average rank correlation across a pre-sample epoch; navy, average rank correlation across sample epoch; red, average rank correlation across delay epoch; green, average rank correlation across response epoch; cyan, average rank correlation between pre-sample and sample epochs; yellow, average rank correlation between sample and delay epochs; magenta, average rank correlation between delay and response epochs). Rank correlation is higher in shared-input space ($p < .001$, paired t-test), which indicates that it is not simply the temporally uncorrelated noise that disrupts the continuity of neural dynamics of trial identity, but some fundamental independent variables affecting each neuron. **d.** We examined whether any

systematic bias drives a higher rank correlation of trial identity in a specific trial type. Across sessions, our study showed equal drive for both trial types (a bias could only exist in a small session with a few units; p = .468, paired t-test).

Figure S4. (Supporting figure for Figures 3.1, 3.5) Rank correlation of trials using neural dynamics in Gaussian Process Factor Analysis space

We compared the same analysis in **Figure 3.5** based on a TLDS model with that from a popular latent-space analysis model, named Gaussian Process Factor Analysis (GPFA) (Yu et al., 2009). **a.** Same as **Figure 3.1c** using the GPFA model. **b.** Same as **Figures 3.5a-d** using GPFA model. **c.** The same comparison as **Figure 3.S3c** between neural dynamics in shared-activity space (TLDS model) and that in GPFA space (GPFA model). Neural dynamics in shared-activity space exhibited a better separation across trial types (left, $p < .001$, paired t-test), while the separation of trial identity in the same trial type was similar in shared-activity space to that in GPFA space. **d.** The optimal dimension of GPFA space increases, significantly in a given range, as a function of the number of simultaneously recorded units ($r_s = .80$, $p < .001$) (each circle represents a simultaneously recorded session; color, optimal dimension of shared-activity space; navy, the optimal dimension of shared-input space is 2; blue, 3; green, 4; yellow, 5). **e.** The explained variance of neural dynamics is similar in shared-input space to that in GPFA space (p = .532, paired t-test; each

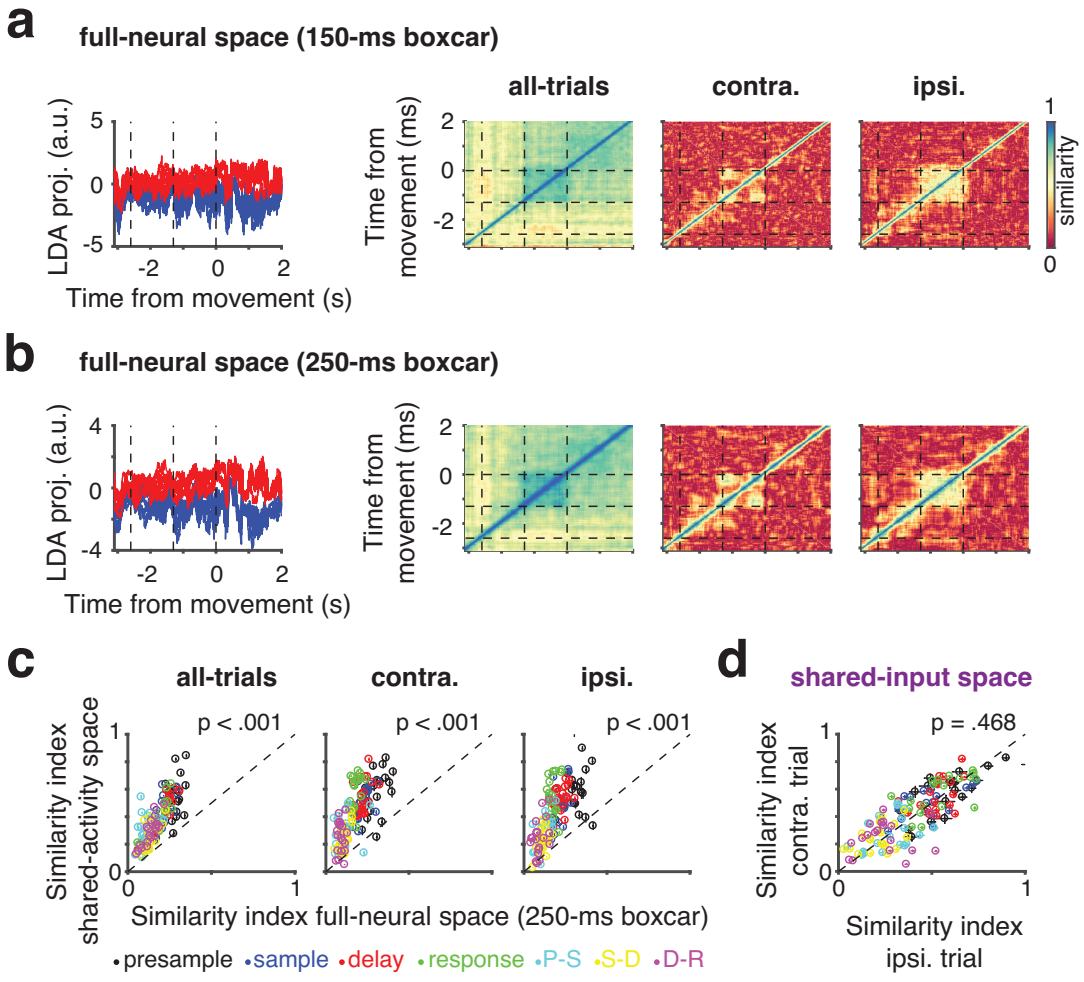


Figure 3.S3: Rank correlation of trials is higher in shared-input space than that in boxcar-smoothed full-neural space.

circle represents a simultaneously recorded session).

Figure S5. (Supporting figure for Figure 3.2d) Details of rank correlation between neural-mode LDA score and reaction time

Estimation of rank correlation between the neural-mode LDA score and reaction time could be influenced if the range of behavioral variability to reaction time was small. We examined this effect by plotting the rank correlation against the standard deviation of behavioral reaction time in each trial type. **a.** Plot of the rank correlation in shared-input space against the standard deviation of behavioral reaction time in contra. trial (significant correlation, filled circle, $p < .05$; otherwise, empty circle; color follows the same schematic as **Figure 3.2d**). **b**. Same as **a** for ipsi. trial. **c-d**. The same as **a-b** in full neural space, where the neural activity was the averaged in a 350 ms time window before response. **e.** Same as **Figure 3.2d** in full neural space, where the neural activity was the averaged in a 350 ms time window before response.

3.5.2 Supplementary tables

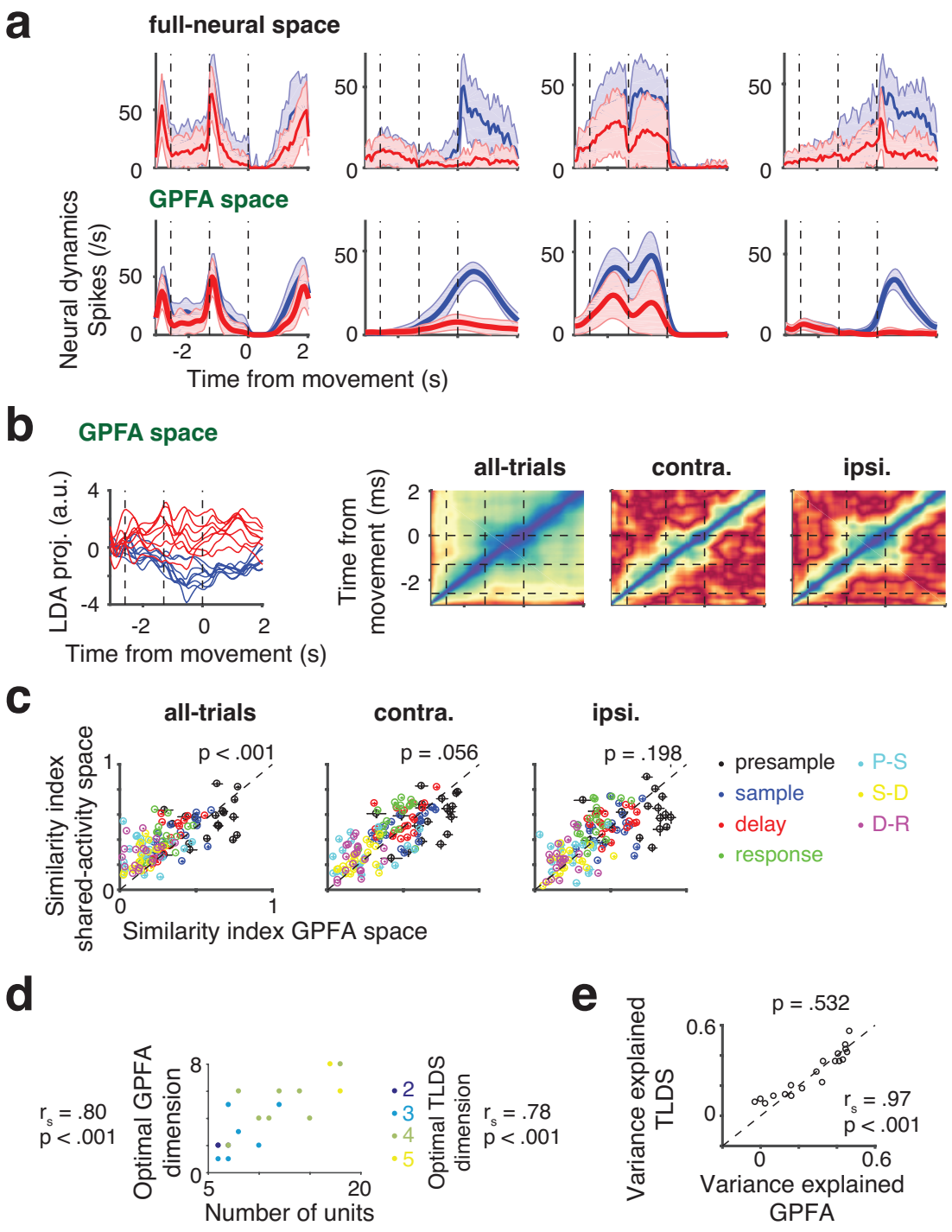


Figure 3.S4: Rank correlation of trials using neural dynamics in Gaussian Process Factor Analysis space.

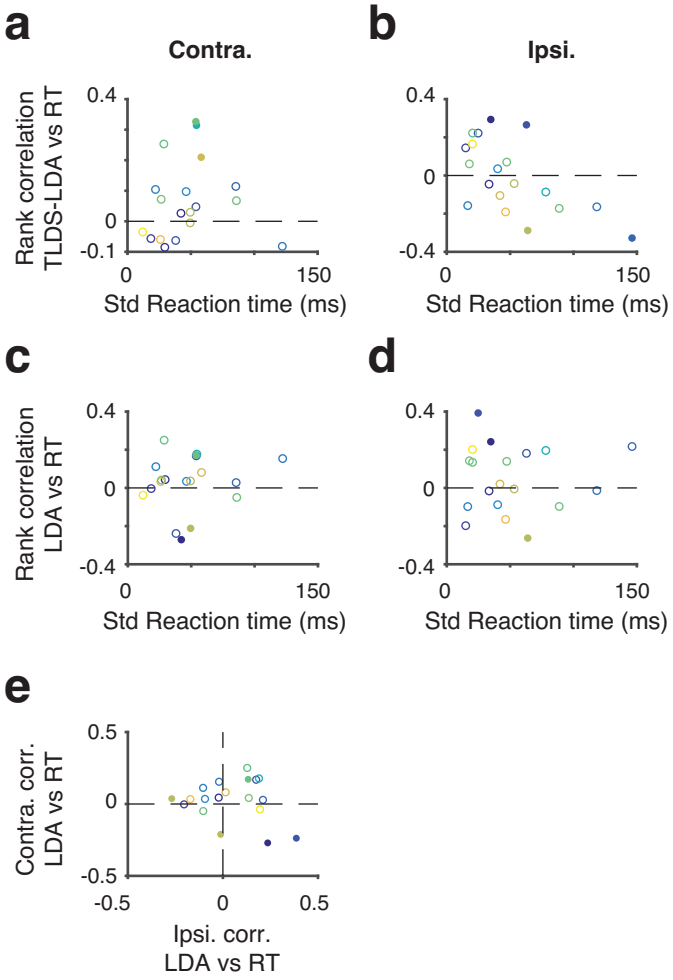


Figure 3.S5: Rank correlation between neural-mode LDA score and reaction time.

Table 3.S1: Correct trial information

		Contra. Correct		Ipsi. Correct	
Session Index	# units	# trials	Spike range (Hz)	# trials	Spike range (Hz)
#01	6	23	4.47 - 13.16	24	5.32 - 11.44
#02	7	45	5.89 - 25.25	29	4.83 - 27.83
#03	8	32	3.11 - 25.08	29	3.46 - 24.76
#04	7	22	4.06 - 19.47	22	4.17 - 18.32
#05	11	34	9.31 - 40.76	33	2.66 - 35.01
#06	6	39	5.59 - 15.59	53	5.51 - 17.73
#07	7	38	5.74 - 23.21	18	3.46 - 26.20
#08	6	38	4.33 - 32.43	33	4.00 - 28.35
#09	7	36	0.65 - 12.80	49	0.38 - 12.58
#10	8	76	1.10 - 12.74	23	0.53 - 11.67
#11	15	51	0.79 - 24.97	67	0.92 - 17.22
#12	10	40	0.47 - 24.99	33	0.50 - 26.99
#13	12	36	0.60 - 18.88	32	0.62 - 17.19
#14	17	133	0.87 - 18.86	98	0.71 - 15.24
#15	18	84	0.87 - 22.75	59	0.45 - 18.09
#16	10	82	0.68 - 20.25	80	0.48 - 18.27
#17	12	117	0.90 - 35.06	70	0.55 - 31.76
#18	14	115	2.32 - 29.42	94	1.94 - 23.15
#19	18	62	0.91 - 30.92	53	1.37 - 23.62

Table 3.S2: Error trial information

		Contra. Error		Ipsi. Error	
Session Index	# units	# trials	Spike range (Hz)	# trials	Spike range (Hz)
#01	6	4	5.20 - 16.86	5	4.09 - 15.61
#02	7	15	5.72 - 26.03	6	4.53 - 25.02
#03	8	8	3.54 - 24.12	3	4.18 - 28.72
#04	7	7	4.21 - 20.07	5	3.31 - 17.46
#05	11	7	4.54 - 41.46	4	7.81 - 41.77
#06	6	20	4.99 - 18.85	5	4.47 - 18.54
#07	7	7	2.78 - 23.53	29	4.07 - 23.53
#08	6	9	4.20 - 31.35	32	4.67 - 31.35
#09	7	5	0.42 - 13.80	19	0.38 - 10.81
#10	8	15	0.63 - 16.50	35	0.80 - 14.87
#11	15	13	1.04 - 18.43	26	0.98 - 20.70
#12	10	9	0.41 - 25.12	10	0.40 - 22.92
#13	12	3	0.45 - 17.09	10	0.83 - 19.12
#14	17	10	0.56 - 15.88	11	0.91 - 14.42
#15	18	27	0.58 - 12.66	7	0.44 - 18.89
#16	10	7	0.36 - 13.93	12	0.21 - 17.17
#17	12	2	0.48 - 31.41	4	0.24 - 35.32
#18	14	19	0.84 - 27.60	5	1.43 - 32.42
#19	18	14	1.07 - 27.62	12	0.53 - 32.12

Table 3.S3: Recording methods, depth and cell type information

			Depth range (um)		Cell type	
Session Index	# units	Recording	min	max	# pyr	# int
#01	6	32-channel	274.04	701.10	4	2
#02	7	32-channel	390.41	710.71	5	2
#03	8	32-channel	370.13	903.96	7	1
#04	7	32-channel	293.25	933.85	6	1
#05	11	32-channel	371.19	905.03	8	3
#06	6	32-channel	489.70	810.00	4	2
#07	7	32-channel	230.26	657.33	5	2
#08	6	32-channel	577.25	897.55	6	0
#09	7	64-channel	219.96	849.83	6	1
#10	8	64-channel	265.48	670.77	7	1
#11	15	64-channel	297.17	862.47	15	0
#12	10	64-channel	369.47	817.66	8	2
#13	12	64-channel	434.48	950.40	10	2
#14	17	64-channel	573.06	1008.77	14	3
#15	18	64-channel	466.89	977.70	13	5
#16	10	64-channel	748.58	1004.86	7	3
#17	12	64-channel	442.59	1043.90	9	3
#18	14	64-channel	489.18	1030.59	9	5
#19	18	64-channel	364.20	1009.08	13	5

Table 3.S4: Explained variance

Session Index	Correct		Error		Correct mean model
	Mean	Sem	Mean	Sem	Mean
#01	.09	.02	.00	.08	.33
#02	.22	.02	.16	.04	.42
#03	.13	.01	.00	.08	.43
#04	.20	.02	.10	.05	.48
#05	.36	.02	.24	.06	.54
#06	.08	.01	.02	.03	.43
#07	.11	.02	.00	.04	.42
#08	.56	.02	.22	.04	.55
#09	.13	.04	.09	.09	.25
#10	.14	.03	-.03	.07	.38
#11	.42	.02	.22	.02	.54
#12	.18	.02	.00	.03	.51
#13	.29	.01	.18	.05	.47
#14	.36	.01	.22	.03	.48
#15	.47	.01	.27	.02	.57
#16	.41	.01	.32	.03	.49
#17	.37	.01	.33	.04	.48
#18	.44	.01	.49	.01	.50
#19	.36	.01	.25	.02	.50

Table 3.S5: Neural dynamics and reaction time correlation

	LDA-RT correlation				TLDS LDA-RT correlation			
Session Index	Contra		Ipsi		Contra		Ipsi	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
#01	.04	.35	-.02	.34	-.09	.35	-.05	.34
#02	.03	.25	-.09	.31	.10	.25	.03	.31
#03	-.24	.28	**.39**	.27	-.06	.30	.22	.30
#04	.11	.36	-.10	.36	.10	.36	-.16	.35
#05	.25	.27	.13	.29	.25	.27	.22	.28
#06	**-.27**	.25	**.24**	.22	.02	.27	**.29**	.21
#07	-.01	.27	-.20	.39	-.06	.27	.14	.39
#08	-.04	.27	.20	.28	-.04	.27	.16	.28
#09	.16	.27	.18	.23	.05	.28	**.26**	.22
#10	.02	.19	.21	.34	.11	.19	**-.33**	.32
#11	.03	.23	**-.27**	.19	-.01	.23	**-.29**	.19
#12	.15	.26	-.02	.29	-.08	.26	-.17	.28
#13	.17	.27	.19	.29	**.31**	.25	-.09	.29
#14	**.17**	.14	.14	.16	**.32**	.13	.07	.17
#15	.03	.18	-.17	.21	-.06	.18	-.19	.21
#16	**-.22**	.17	-.01	.19	.03	.18	-.05	.18
#17	.04	.15	.14	.19	.07	.15	.06	.20
#18	.08	.15	.02	.17	**.21**	.15	-.11	.17
#19	-.05	.21	-.10	.23	.07	.21	-.18	.22

Table 3.S6: Effective eigenvalues

Session Index	Sample	Delay	Response
#01	0.98	0.98	0.97
#02	0.98	1.00	0.98
#03	1.01	1.00	1.00
#04	0.93	0.92	1.00
#05	1.01	0.99	0.94
#06	0.85	0.99	0.89
#07	0.98	0.99	1.00
#08	0.98	0.99	1.00
#09	0.90	0.92	0.96
#10	0.99	0.98	0.99
#11	0.97	1.00	1.01
#12	0.96	0.94	0.96
#13	0.99	0.99	0.94
#14	1.00	1.00	0.99
#15	1.00	0.99	1.00
#16	0.98	1.01	0.99
#17	0.97	1.01	1.01
#18	1.00	0.99	1.00
#19	1.00	1.00	1.02

Chapter 4

Confidence Estimation as a Stochastic Process in a Neural Dynamical System of Decision Making

Evaluation of confidence about one's knowledge is key to the brain's ability to monitor cognition. To investigate the neural mechanism of confidence assessment, we examined a biologically realistic spiking network model and found that it reproduced salient behavioral observations and single-neuron activity data from a monkey experiment designed to study confidence about a decision under uncertainty. Interestingly, the model predicts that changes of mind can occur in a mnemonic delay when

confidence is low; the probability of changes of mind increases (decreases) with task difficulty in correct (error) trials. Furthermore, a so-called "hard-easy effect" observed in humans naturally emerges, i.e. behavior shows under-confidence (underestimation of correct rate) for easy or moderately difficult tasks, and overconfidence (overestimation of correct rate) for very difficult tasks. Importantly, in the model, confidence is computed using a simple neural signal in individual trials, without explicit representation of probability functions. Therefore, even a concept of metacognition can be explained by sampling a stochastic neural activity pattern (Wei and Wang, 2015).

4.1 Introduction

A key to the monitoring of cognition (metacognition) is our ability to evaluate the degree of confidence that we have about a decision, a strategy to tackle the problem at hand, a newly acquired piece of knowledge, and so on. Confidence estimation has been an important subject of research in cognitive and developmental psychology (Flavell, 1979; Vickers, 1979). In laboratory studies, confidence can be measured using post-decision wagering (PDW), where subjects first perform a perceptual decision, and then make a high-low bet between a risky option (associated with a high reward if the first-order choice is correct, a loss otherwise) and a safe option (associated with a low reward regardless of the first-order choice). If subjects have less confidence about their choice, they should be more likely to bet on the low but certain reward option

(Persaud et al., 2007; Dienes and Seth, 2010; Fleming and Dolan, 2010; Fleming et al., 2010; Kepecs and Mainen, 2012). Recently, researchers have begun to use PDW with behaving animals to explore the neural basis of confidence estimation (Smith et al., 2003; Kepecs et al., 2008; Middlebrooks and Sommer, 2011; Middlebrooks and Sommer, 2012; Lak et al., 2014).

In a monkey experiment, Kiani and Shadlen (2009) extended a well-known discrimination task to examine neural signals correlated with confidence. In this task, a subject is required to decide between two possible directions (indicated by two directional targets) of a random-dots motion stimulus. Specifically, Kiani and Shadlen used a fixed-duration (FD) version of the task (Shadlen and Newsome, 2001), where the visual motion stimulus is followed by a delay period, and monkeys must indicate the decision at the end of the delay by a saccadic response to one of the directional targets. In a random subset of trials, they offered a third target (namely the "sure" target, T_s) during the delay period, and monkeys could opt to T_s for a certain but small amount of reward. Interestingly, monkeys selected T_s more often when motion strength was weaker or stimulus duration became shorter, under which conditions the error rate was higher and selecting T_s gave rise to an overall increases in rewards across trials. The probability of choosing T_s (P_{sure}) thus reflected a degree of choice uncertainty. Importantly, P_{sure} was found to be correlated with single-neuron activity in the lateral intraparietal (LIP) area, an area that was correlated with accumulating decision evidence of a choice. This finding supports the intuitive idea that confidence

signal is an integral part of a decision process (Vickers, 1979), and reflected in a neural decision variable (Gold and Shadlen, 2007).

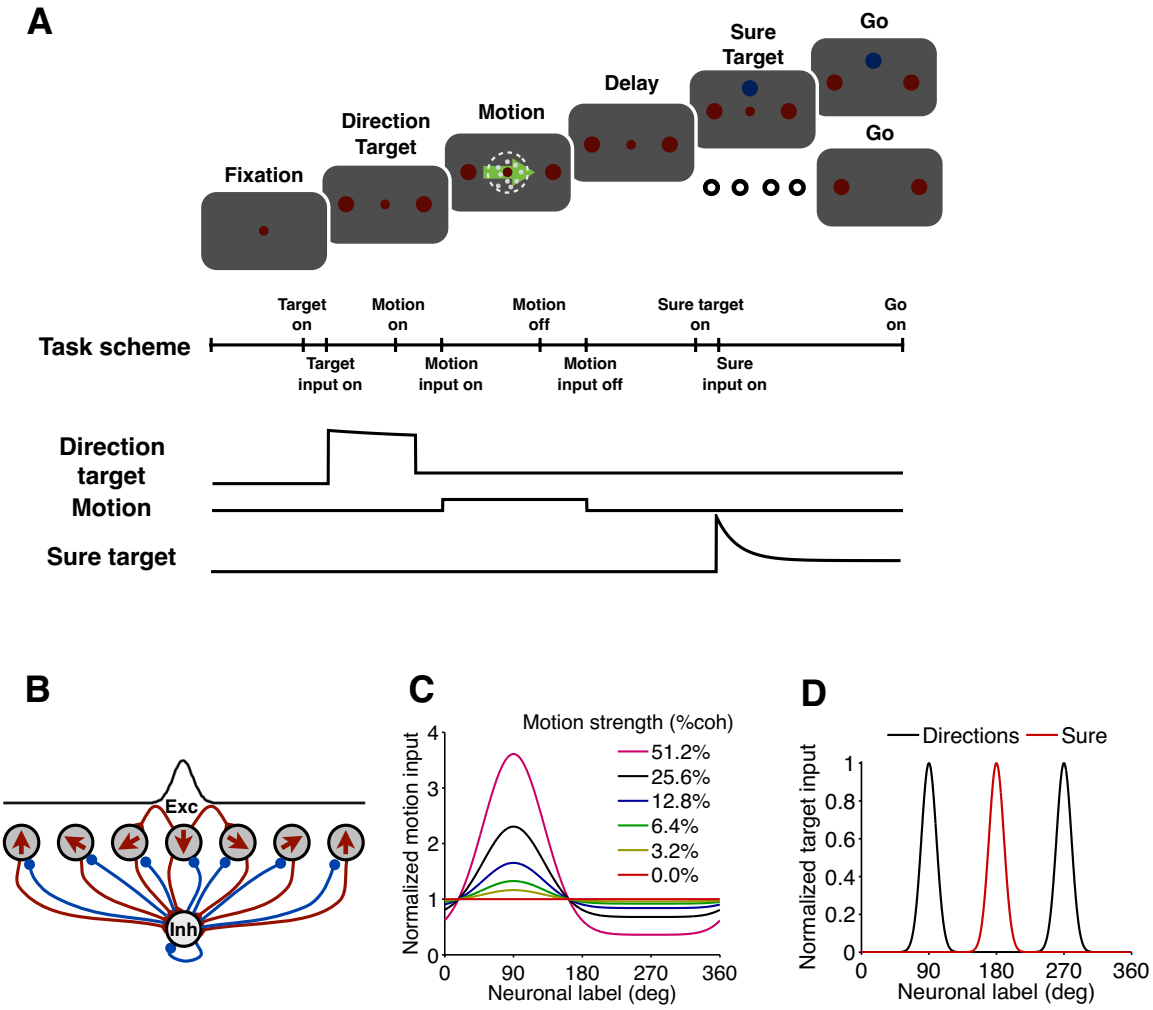
Computational schemes have been proposed for the study of confidence (Vickers, 1979; Kepecs et al., 2008; Ratcliff and Starns, 2009; Kiani and Shadlen, 2009; Pleskac and Busemeyer, 2010; Rolls et al., 2010a; Rolls et al., 2010b; Moreno-Bote, 2010; Kepecs and Mainen, 2012). In particular, using drift diffusion model (DDM) (Ratcliff and Smith, 2004), confidence has been defined in terms of the log posterior ratio for the two choices given the decision variable (DV) at the time of behavioral response (Kiani and Shadlen, 2009). This DDM, nevertheless, has some limitations when accounting for the complexity of confidence estimation, e.g. a DV that terminates at a fixed threshold may not present a graded confidence across trials (Kiani et al., 2014), and the behavioral change of receiver operating characteristic (ROC) curve under bias cannot be explained (Van Zandt, 2000).

In this work, in order to uncover the neural circuit mechanism underlying confidence estimation, we took a different approach and employed a biophysically realistic cortical network model of spiking neurons, which was previously shown to successfully simulate the two-target random-dots motion direction discrimination experiment (Furman and Wang, 2008). We investigated whether the same model could accurately reproduce the salient findings from Kiani and Shadlen (2009). The model is endowed with a continuous network of neurons that can represent any direction; therefore it can be readily extended to incorporate the presentation of a sure target (T_s) during

a delay period. Notably, such a model of decision-making and memory processes was not originally designed for the Kiani-Shadlen experiment to account for confidence estimation. It is thus surprising that the model can capture both a range of behavioral performance data and physiological observations from LIP single neurons. We noted that neurons selective for T_s win the competition (thus T_s was chosen) when the activities of neurons selective for the two alternative choices are indistinguishable. Quantitatively, we found that confidence could be estimated, at any time, as a sigmoid function of the differential firing activity of the two competing neural pools selective for the alternative choices (Beck et al., 2008). Therefore, choice confidence is computed simultaneously when a decision is made, and a trial-by-trial variation of choice is generated by sampling of stochastic neural dynamics (Wang, 2008).

4.2 Results

We performed computer simulations of Kiani-Shadlen task (**Figure 4.1A**), using a neuronal decision model (Furman and Wang, 2008). In this model, the pyramidal cells are selective for motion direction as an analog stimulus feature, and are uniformly distributed along a ring according to their preferred directions. Pyramidal cells are endowed with strong recurrent excitation, which is balanced by feedback inhibition mediated by interneurons (**Figure 4.1B**). We assumed that the neural representation of motion stimulus in middle-temporal area (MT) exhibits normalization (Heeger,


A
Fixation
Direction Target
Motion
Delay
Sure Target
Go
Go
Task scheme
Target on
Motion on
Motion off
Sure target on
Go on
Target input on
Motion input on
Motion input off
Sure input on
Direction target
Motion
Sure target
B
Exc
Inh
C
Motion strength (%coh)
51.2%
25.6%
12.8%
6.4%
3.2%
0.0%
Normalized motion input
Neuronal label (deg)
D
Directions
Sure
Normalized target input
Neuronal label (deg)

1992; Treue et al., 2000) (**Figure 4.1C**). The output from MT converges with other visual inputs such as choice targets to the decision circuit. Without loss of generality, we placed the choice targets T_A at 90°, T_B at 270°, and T_s at 180° (**Figure 4.1D**). In a short delay after the target onset, a random-dots motion stimulus is presented and, in our model, the network integrates the motion signal gradually over time. If a categorical choice is formed through attractor dynamics (Wang, 2002; Wong and Wang, 2006; Furman and Wang, 2008; Wang, 2008), that choice is maintained in the form of a persistent activity pattern during a delay period. Since the network dynamics are stochastic, a decision may not be reached during stimulus presentation. Network activity continues to evolve during the delay through slow NMDA-mediated

Figure 4.1 *(preceding page)*: Schematic description of the decision task and model architecture
(**A**) Procedure of a simulated fixed-duration discrimination task. Following a fixation period, two targets (large red circles) appear, indicating the alternative choices. A random-dots motion stimulus is presented for 110 ms to 627 ms followed by a delay period. A saccade to one of the alternatives indicates the decision at the end of the delay. In some trials, a sure target (blue circle) is shown from 675 ms after the motion offset, and choosing it leads to a certain but small amount of reward. A detailed task scheme is shown in lower panel; the input of target (motion, and sure target, respectively) is delayed 100 ms (200 ms, and 100 ms, respectively) to the onset of target (motion, and sure target, respectively). (**B**) Neural network structure. The network consists of excitatory pyramidal cells (EXC) and inhibitory interneurons (INH). The pyramidal cells are uniformly placed on a continuous ring and each neuron is labeled by its preferred motion direction (shown as the arrow in pyramidal cells). The excitatory-to-excitatory connections between pyramidal cells are structured as a Gaussian function of the difference in their preferred directions (upper black curve) and the connections from and onto interneurons are broad. (**C**) Motion input (centered at 90°) with different motion strengths, the integral of which is identical for all motion strengths. (**D**) Normalized input of two directional targets (namely T_A at 90° and T_B at 270°), and a sure target (namely T_s at 180°).

reverberation, and this process can be altered in the event that T_s is presented and this third option becomes available.

4.2.1 Network dynamics in a fixed duration task with post decision wagering at zero motion strength

In experiments using single-unit recording, each neuron was recorded independently, one at a time, and its selectivity and dynamics were evaluated across trials, whereas in our model, all neurons are monitored simultaneously in a single trial. At the population level, the ramping activity is demonstrated as the gradual development of a bell-shaped activity pattern (bump) around the direction of a selected target. The stimulated neural dynamics in **Figure 4.2** can be compared directly with single-neuron data from area LIP: for a neuron with the preferred direction at T_A, T_A and T_B are equivalent to T_{in} and T_{opp} in Kiani-Shadlen experiment (2009).

Figure 4.2A shows the spatiotemporal spiking activity of the network model in a trial without T_s presented. Although the input is identical to all the pyramidal cells at zero motion strength, the two activity bumps compete with each other through shared inhibitory feedback and stochastic recurrent dynamics. Eventually one bump ramps up, while the other one decays, leading to a categorical choice (Wang, 2002; Wong et al., 2007; Furman and Wang, 2008). The ramping-up bump is maintained

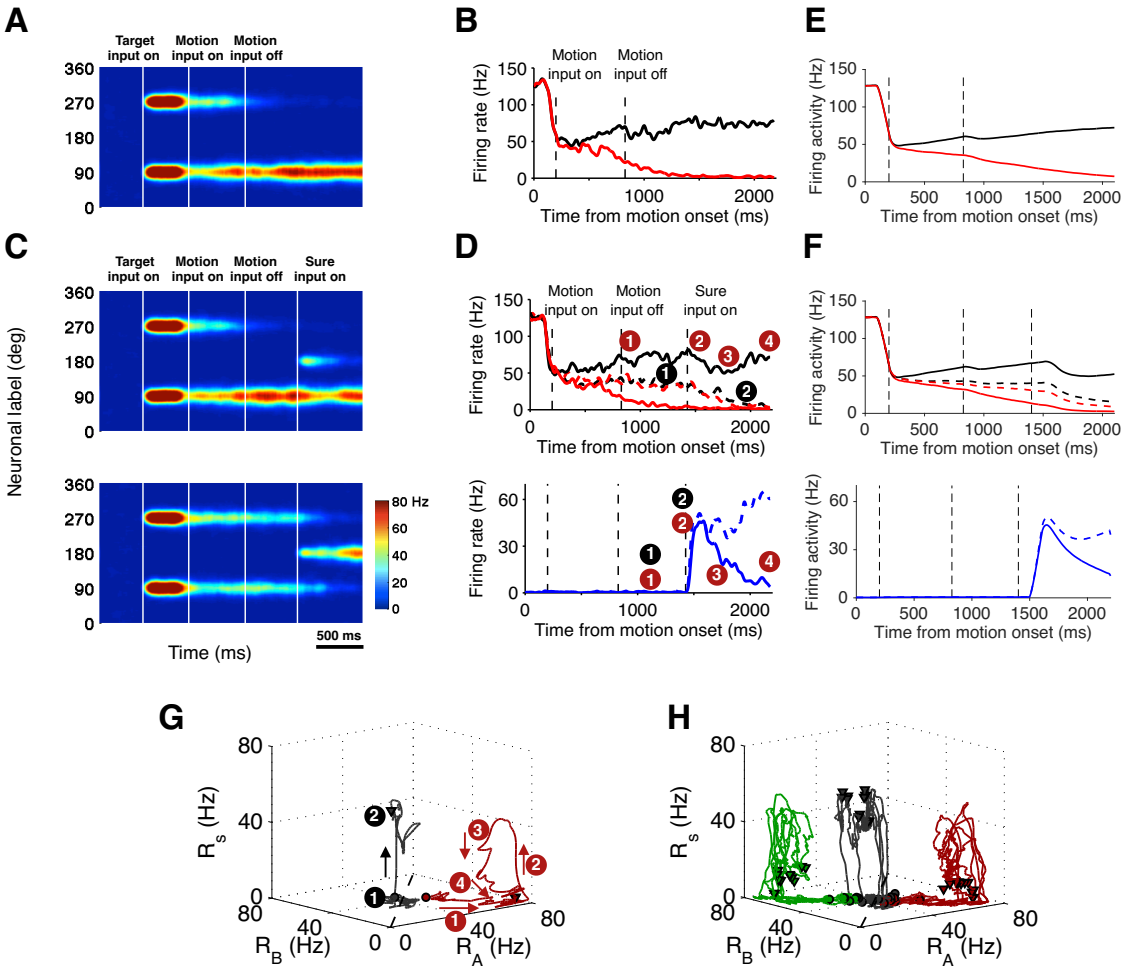

A
Target input on
Motion input on
Motion input off
Sure input on
Neuronal label (deg)
360
270
180
90
0
80 Hz
60
40
20
0
500 ms
Time (ms)
B
Firing rate (Hz)
150
100
50
0
1000
2000
Time from motion onset (ms)
C
D
E
Firing activity (Hz)
500
1500
F
60
30
G
H
R_S (Hz)
R_B (Hz)
R_A (Hz)
80
40

Figure 4.2 *(preceding page)*: Neuronal activity of sample trials at zero motion strength
(**A**) A sample trial where T_s was not presented (stimulus duration is 627 ms). Neural pools centered around the two directional targets eventually diverge from each other, that near T_A wins the competition and its activity persists during the delay in the form of a bell-shaped "bump attractor". (**B**) Average firing rate of the neural pools at T_A (black line) and T_B (red line) of the trial in (**A**). (**C**) Two sample trials where T_s was presented (stimulus duration is 627 ms). Upper panel: the sure target induced a transient response which was suppressed due to feedback inhibition within the circuit and the neural pool at T_A preserves similar activity to that in (A), therefore T_s was waived. Lower panel: the neural pool around T_s fires at a sufficiently high rate that it overcomes the competition with the other neural pools which in turn is suppressed by feedback inhibition, therefore T_s was selected. Note that in this trial the neural activities of two competing bumps are indistinguishable prior to T_s onset, and gradually decay to a low level after T_s onset. (**D**) Neural activities at T_A (black lines), T_B (red lines), T_s (blue lines) of the trials in (**C**). Dashed curves: the trial where T_s was selected; solid curves: the trial where T_s was waived. Note that the stimulus condition was identical for the two sample trials, whether the sure target was chosen or waived was completely determined by network dynamics that fluctuated from trial to trial. (**E-F**) Average activities of R_A, R_B and R_s across different motion strengths, which follow the same conventions as those in Figures B and D. (**G-H**) Network dynamics underlying a trial-by-trial variation of choice in a 3D (R_A, R_B, R_s) decision state space: (**G**) neural activity trajectories of the two sampling trials in (C) from 150 ms after motion onset, the starting (end, respectively) points of which are marked by circles (triangles, respectively); time sequence of the trial waiving T_s (black line, 1-2 steps) follows that network walks around R_A=R_B before T_s onset (Step 1 in 2**D**,**E**; black circles), and then converges to R_s after T_s onset (Step 2 in 2**D**,**E**; black circles); time sequence of the trial selecting T_s (red line, 1-4 steps) follows that network goes towards to R_A before T_s onset (Step 1 in 2**D**,**E**; red circles), and then walks along R_s direction after T_s onset (step 2 in 2**D**,**E**; red circles), then converges back to R_A again (Step 3-4 in 2**D**,**E**; red circles); (**H**) those of the other 18 sampling trials at the same stimulus condition. For the trials waiving T_s, the network first converges to a choice attractor T_A (red lines) or T_B (green lines) preceding T_s onset; it then moves along the direction parallel to R_s axis due to the presentation of T_s, and finally converges back to the initial choice attractor. While for the trials choosing T_s (grey lines), the network first walks around the diagonal line R_A=R_B ($R_s \simeq 0$) and then converges to the sure attractor T_s after its presentation. Neural dynamics therefore acts as a 3-way competition and the basins of attraction are clearly separated.

by the persistent activity in delay (Compte et al., 2000). At the single-unit level, the firing activities of neurons at T_A and T_B (R_A and R_B, respectively) diverge over time after motion onset leading to the network's choice (T_A). The winning neural pool persists its firing activity till the end of delay (**Figure 4.2B**).

In trials when T_s is presented (**Figures 4.2C-D**), again the attractor dynamics dictate the network's choice behavior, and we identified a choice of T_s by the firing activity of neurons at T_s (R_s): the network selects T_s (**Figure 4.2D**, solid blue line), if R_s persists at a high rate; it waives T_s (**Figure 4.2D**, dash blue line), if R_s decays to a low rate. In the trial where T_s is waived, we observed the same divergence and persistent activity of R_A (choice) and R_B as that without T_s presented (**Figure 4.2C**, upper panel; **Figure 4.2D**, solid lines); however, if the network selects T_s, R_A and R_B are indistinguishable at some intermediate rates without significant divergence (**Figure 4.2C**, lower panel; **Figures 4.2D**, dash lines; R_A=32.5 Hz, R_B=33.6 Hz at T_s onset), and then both decay to low rates as neurons around T_s win the competition. The average firing activities of R_A, R_B, and R_s across different motion strengths (100 trials for each motion strength) are shown in **Figures 4.2E-F**, which follow the same conventions as those in **Figure 4.2B** and **Figure 4.2D**. All these are similar to the observed LIP neuronal data (Kiani and Shadlen, 2009) from motion input onset time to the end of trial, which is the critical period for capturing the neural dynamics of choosing and waiving T_s in the experiment. While a difference of neuronal activities could exist before motion input onset in our model, compared to the experimental

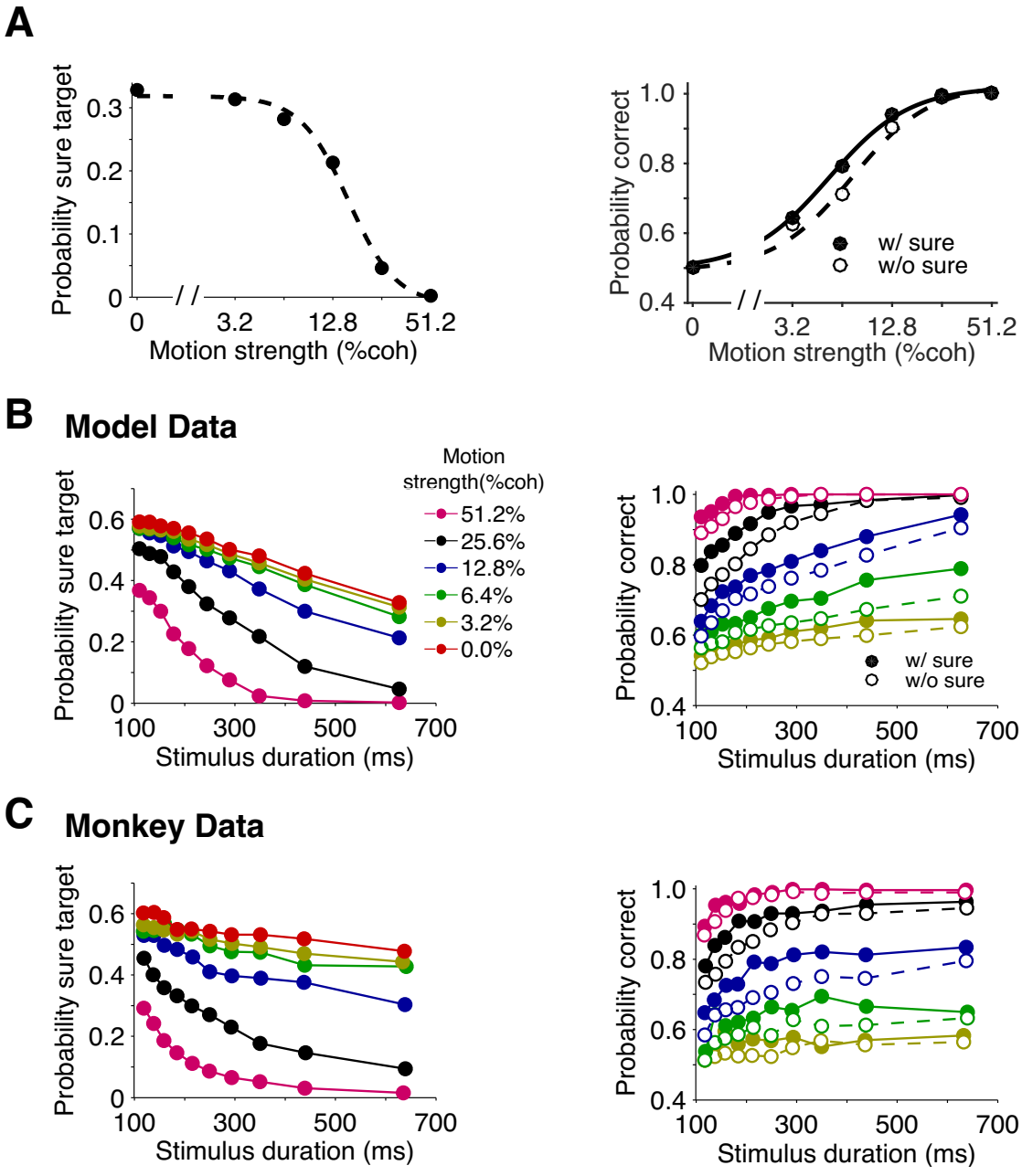
observations, this is not important for predicting a choice of T_A, T_B or T_s.

We further studied the neural dynamics that underlies a choice of the network among T_A, T_B and T_s. Taking trials in **Figure 4.2C** for example, we visualized the neural trajectories in a 3D decision space (R_A, R_B, R_s) following the sequences marked in Figure 2D (**Figure 4.2G**). In the trial where T_s is waived (**Figure 4.2G**, red line), the network first converges to a choice attractor (T_A), then leaves away from and returns to it again after the presentation of T_s; in the trial where T_s is selected (**Figure 4.2G**, grey line), the network wanders around the diagonal line (R_A=R_B, $R_s \simeq 0$) and then converges to the attractor T_s. These trajectories in decision space imply that the presentation of T_s could induce a sure attractor, which behaves similar to choice attractors; the network could thus act like a three-way competition after the presentation of T_s. To test this, we visualized more sampling trials at the same stimulus condition and explored the basin of attraction for each attractor (**Figure 4.2H**). We found that for trials choosing T_A, T_B and T_s (**Figure 4.2H**, red, green and grey lines, respectively), the networks converge to the choice attractors T_A (near the R_A axis), T_B (near the R_B axis) and the sure attractor T_s (near the R_s axis), respectively. The whole decision space is thereby separated into three attractor regions, and due to this structure of decision space, the location of each neural trajectory at the moment of T_s onset (namely the initial location) can potentially predict the choice of a network (Kiani and Shadlen, 2009). Particularly, if the initial location of a network is close to a choice attractor, it would eventually

converge back to that choice attractor again after T_s onset, while if its initial location is around the diagonal line, it more likely converges to a sure attractor. Notably, if the initial location of a network is between the diagonal line and a choice attractor, it could either continue converging to that choice attractor or change its mind to T_s (We will discuss it later in **Figure 4.7**). Therefore, our model demonstrates that a categorical choice of a network in this task could be generated by a 3-way competition among attractors T_A, T_B and T_s, which relies internally on the stochastic neural dynamics (Wang, 2008).

4.2.2 Behavioral performance

The model's performance is quantified by the fraction of trials corresponding to a particular behavioral response. **Figure 4.3** exhibits the probability of choosing T_s (P_{sure}) and accuracy ($P_{correct}$) for trials when T_s is not presented or T_s is shown but waived. At a fixed stimulus duration, our model shows that P_{sure} decreases while $P_{correct}$ increases with the motion strength; $P_{correct}$ improves in trials where T_s was shown but waived (**Figure 4.3A**). Moreover, **Figure 4.3B** shows that P_{sure} decreases with the stimulus duration, therefore the network selects T_s more often for weaker motion strength or shorter duration; $P_{correct}$ increases monotonically with stimulus duration for trials with or without T_s presented (**Figure 4.3B**, right panel, solid and dash lines, respectively). For a given motion strength and stimulus duration, $P_{correct}$ is higher for trials where T_s is shown but waived than those without T_s presented,


A
Probability sure target
0.3
0.2
0.1
0
0
3.2
12.8
51.2
Motion strength (%coh)
Probability correct
1.0
0.8
0.6
0.4
w/ sure
w/o sure
0
3.2
12.8
51.2
Motion strength (%coh)
B
Model Data
Motion strength(%coh)
51.2%
25.6%
12.8%
6.4%
3.2%
0.0%
Probability sure target
0.6
0.4
0.2
0
100
300
500
700
Stimulus duration (ms)
Probability correct
1.0
0.8
0.6
0.4
w/ sure
w/o sure
100
300
500
700
Stimulus duration (ms)
C
Monkey Data
Probability sure target
0.6
0.4
0.2
0
100
300
500
700
Stimulus duration (ms)
Probability correct
1.0
0.8
0.6
0.4
100
300
500
700
Stimulus duration (ms)

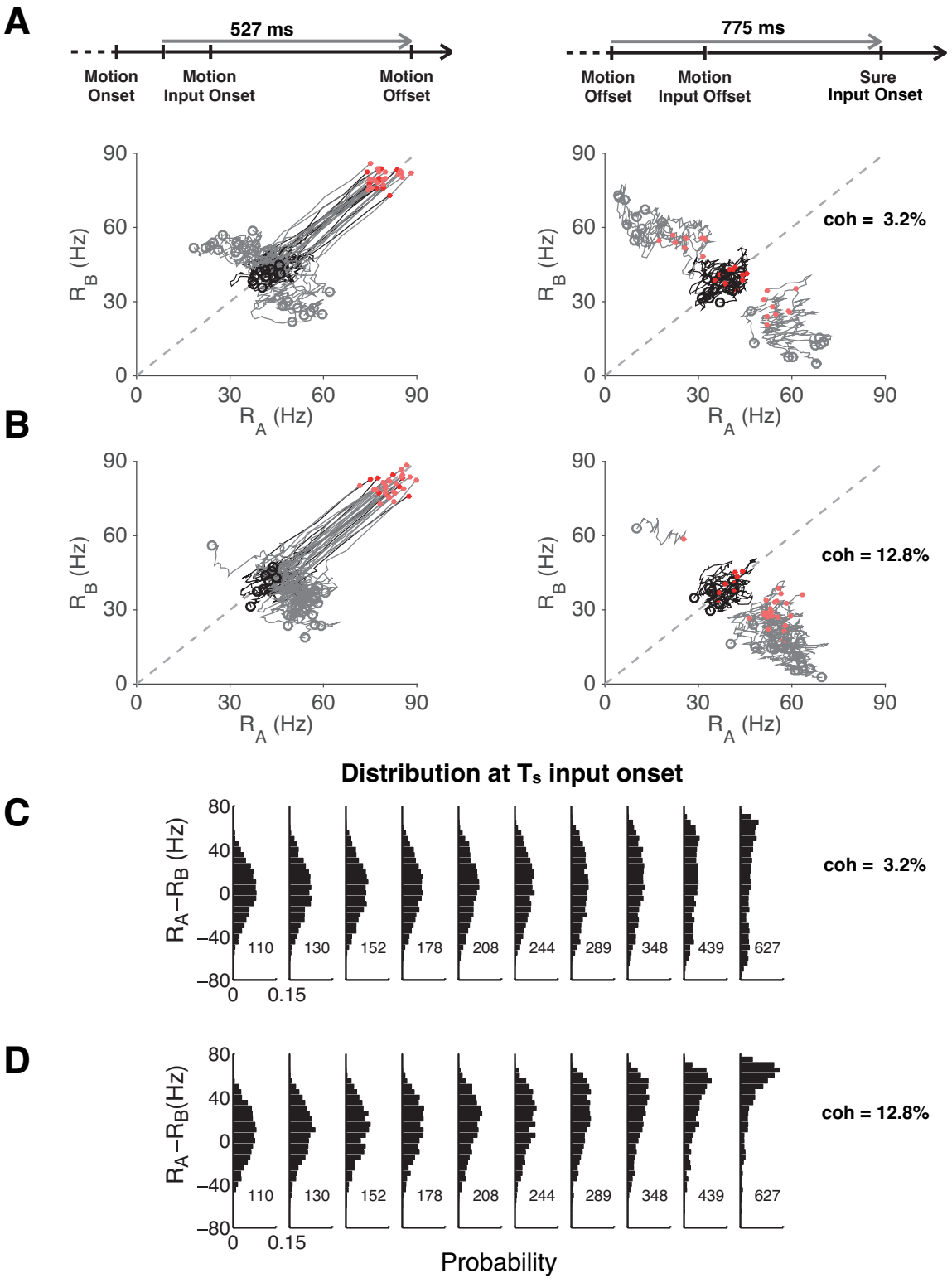
implying that P_{sure} is a probe of uncertainty (Whiteley and Sahani, 2008; Kepecs and Mainen, 2012). In conclusion, the model successfully reproduces the salient behavioral observations in the monkey experiment (Kiani and Shadlen, 2009) (**Figure 4.3C**).

4.2.3 Choice confidence as a logistic function of the differential activity

Consistent with the neurophysiological observation (Kiani and Shadlen, 2009), R_A and R_B undergo brief decreases after T_s onset in our model (**Figure 4.2C**). This is because T_s stimulates neurons selective for the sure target, and their increased firing activity recruits more feedback inhibition that reduces R_A and R_B in a three-way competition (**Figure 4.2H**). We therefore hypothesized that the network would opt to T_s if it has not converged to a stable attractor for T_A ($R_A >> R_B$) or T_B ($R_A << R_B$); the differential activity $|R_A - R_B|$ at T_s onset determines P_{sure}.

We first observed that R_A and R_B could diverge in the late phase of delay. The

Figure 4.3 *(preceding page)*: Behavioral performance
(**A**) Model performance (at a fixed stimulus duration of 627 ms). Left panel: the probability of choosing T_s (P_{sure}) decreases as a sigmoid function of the motion strength; Right panel: accuracy in trials where T_s is not shown ($P_{correct}$) increases as a sigmoid function of the motion strength (filled black curve), and it is improved in trials when T_s was shown but waived (open black curve). (**B**) At different stimulus durations, P_{sure} decreases with motion strength and stimulus duration; $P_{correct}$ is higher in trials where T_s was shown but waived (solid lines, filled circles) than that where T_s was not shown (dash lines, open circles). (**C**) Behavioral data from Kiani-Shadlen (2009) task using awake monkeys. Comparing Figures **B** with **C**, model reproduces salient behavioral observations from the monkey experiment.


A
527 ms
Motion Onset
Motion Input Onset
Motion Offset
775 ms
Motion Offset
Motion Input Offset
Sure Input Onset
B
R_B (Hz)
R_A (Hz)
coh = 3.2%
coh = 12.8%
C
Distribution at T_s input onset
R_A–R_B (Hz)
110
130
152
178
208
244
289
348
439
627
coh = 3.2%
D
coh = 12.8%
Probability

long divergent time ($> T_s$ onset) implies that there is a quasi-stable state at $R_A = R_B$, around which the network could wander, but eventually the network would escape from it and converge to a stable attractor, generating a choice. We next studied whether the network in a state around R_A=R_B would opt to T_s (e.g. **Figure 4.2H**, grey lines). For individual sample trials, we visualized R_A and R_B against each other on a decision space. **Figures 4.4A-B** show that the activity of the network falls down along a diagonal line and then separates into three groups for choices of T_A, T_B and T_s. In trials where T_s is waived, the network converges to one of two stable attractors, T_A or T_B (**Figures 4.4A-B**, gray lines), while in those where T_s is selected, the network walks randomly around the quasi-stable state at $R_A = R_B \simeq 35$

Figure 4.4 *(preceding page)*: Differential activity of two competing choices determines whether a sure target is waived
(**A-B**) Single-trial dynamics of the network in the decision state space, where the population firing rates R_A and R_B are plotted against each other (the starting point of each network trajectory is marked as a red circle and ending point is marked as an open circle). Grey: trials when T_s was waived; black: trials when T_s was selected. The dynamical trajectories are shown from 100 ms after motion onset to its offset (left panels), then to T_s input onset (right panels), at different motion strengths (**A**: 3.2%; **B**: 12.8%; stimulus duration: 627 ms). At the onset of motion stimulus, both R_A and R_B are high ($\sim$ 90 Hz), near the diagonal line, due to the presentation of directional targets. The population dynamics first decays along the diagonal line, induced by a suppression of target inputs after motion onset. In trials when T_s was waived, the network trajectory converges to one of two target attractors (where R_A is high and R_B is low, or vice versa), whereas in trials when T_s was selected, the population dynamics continues to wonder randomly around the diagonal line. The absolute value of differential activity at T_s onset therefore determines whether T_s is waived. (**C-D**) The distribution of R_A-R_B at T_s onset is a function of the motion strength (**C**: 3.2%; **D**: 12.8%) and stimulus duration (presented in each column), where the percentage of trials around zero decreases with the motion strength and stimulus duration. This explains why P_{sure} decreases with the motion strength and stimulus duration.

Hz (**Figures 4.4A-B**, black lines). In summary, once the network converges to a stable attractor prior to T_s onset, T_s is waived; if it wanders around $R_A = R_B$, the network is likely to opt to T_s.

The studies of the similar attractor models (Wang, 2002; Wong and Wang, 2006; Wong et al., 2007; Furman and Wang, 2008; Wang, 2008) showed that early divergence of R_A and R_B (bias to one attractor) determines the probability of choosing T_A and T_B as a function of the motion strength and stimulus duration. One can thus expect that early divergence would also result in a decrease of P_{sure} as a function of the motion strength and stimulus duration. To examine this, we investigated the distributions of $R_A - R_B$ with different motion strengths at T_s onset. **Figures 4.4C-D** show that the percentage of the trials around $R_A = R_B$ decreases with higher motion strength or longer stimulus duration, resulting in a decrease of P_{sure}.

Although the early divergence plays a dominant role in the network dynamics, network continues to evolve via NMDA-mediated reverberatory dynamics; the slow stochastic dynamics could thus drive the network away from $R_A = R_B$ in the later phase of the delay. Consequently, P_{sure} depends on T_s onset time. **Figure 4.5A** displays the evolution of the distribution of $R_A - R_B$ at different times from motion offset, demonstrating that the slow stochastic dynamics also plays an essential role in the networks behavior. Across trials, our model predicts that P_{sure} decreases with longer T_s onset times (**Figure 4.5B**), because more trials settle down to a stable attractor later in the delay, i.e. the percentage of trials with R_A=R_B decreases with

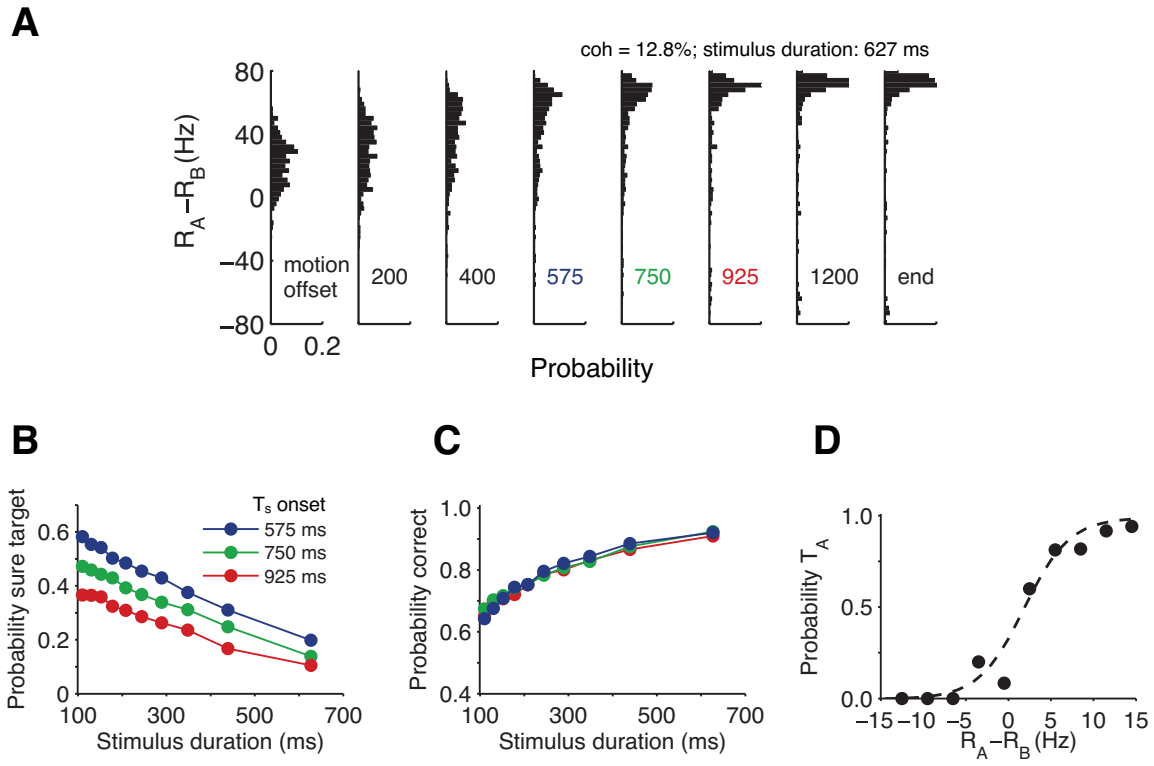


Figure 4.5: Onset time of sure target determines the probability of choosing sure target but has little impact on accuracy

(**A**) At a fixed motion strength and stimulus duration (12.8% and 627 ms, respectively), $R_A - R_B$ continues to change after motion offset (time presented in each column is relative to motion offset), and is settled down only until the late phase of delay (> 1200 ms in simulation). (**B-C**) P_{sure} decreases as a function of T_s input onset time (575 ms: blue; 750 ms: green; 925 ms: red), while $P_{correct}$ remains unaffected. (**D**) Probability of choosing T_A (at the end of the delay) depends on the differential activity, $|R_A - R_B|$, at T_s onset (filled circles: simulation data from **B-C** where coh = 12.8% and motion direction towards to T_A; dashed line: logistic function fit). When $|R_A - R_B|$ is large, the sign of R_A-R_B determines the choice at T_s onset, i.e. positive for T_A, and negative for T_B. If $|R_A - R_B|$ is small ($R_A - R_B$ from -5 Hz to 5 Hz), the probability of choosing T_A increases with $R_A - R_B$. Data in this figure are composed of those at all T_s onset times.

T_s onset. Interestingly, we found that $P_{correct}$ was nearly constant with different T_s onset times (**Figure 4.5C**). This happens because $P_{correct} = \frac{r_{AB}}{1+r_{AB}}$, where r_{AB} is the ratio of the number of trials at attractor T_A (choice) to that at attractor T_B at T_s onset and is saturated at 575 ms after motion offset (**Figure 4.5A**). In short, P_{sure} is directly related to the percentage of trials around R_A=R_B, while $P_{correct}$ is associated with the number of trials at T_A and T_B at T_s onset. In this sense, there is a dissociation of confidence estimation from performance (Graziano et al., 2015; Graziano et al., 2010; Graziano and Sigman, 2009). Moreover, without T_s presented, network continues to converge to T_A or T_B via stochastic dynamics, the probability to one of them is biased, and relies on R_A-R_B in the early phase of delay (**Figure 4.5D**). In conclusion, we found that $|R_A - R_B|$ at the moment of T_s onset determines P_{sure} probabilistically, and reflects a degree of the stability of a choice: if a categorical choice is achieved but with small $|R_A - R_B|$, it could be altered to T_s; whereas if $|R_A - R_B|$ is large, the network's choice would not be changed by T_s.

Here we define choice confidence as a function of the instantaneous differential activity $|R_A - R_B|$ for each trial i, i.e.

$$cc^i = f(|R_A^i - R_B^i|). \tag{4.1}$$

In our model, $|R_A - R_B|$ shows the position of the network state in the (R_A,R_B) plane (**Figures 4.2G-H**) related to choice attractors in the decision space, i.e. the

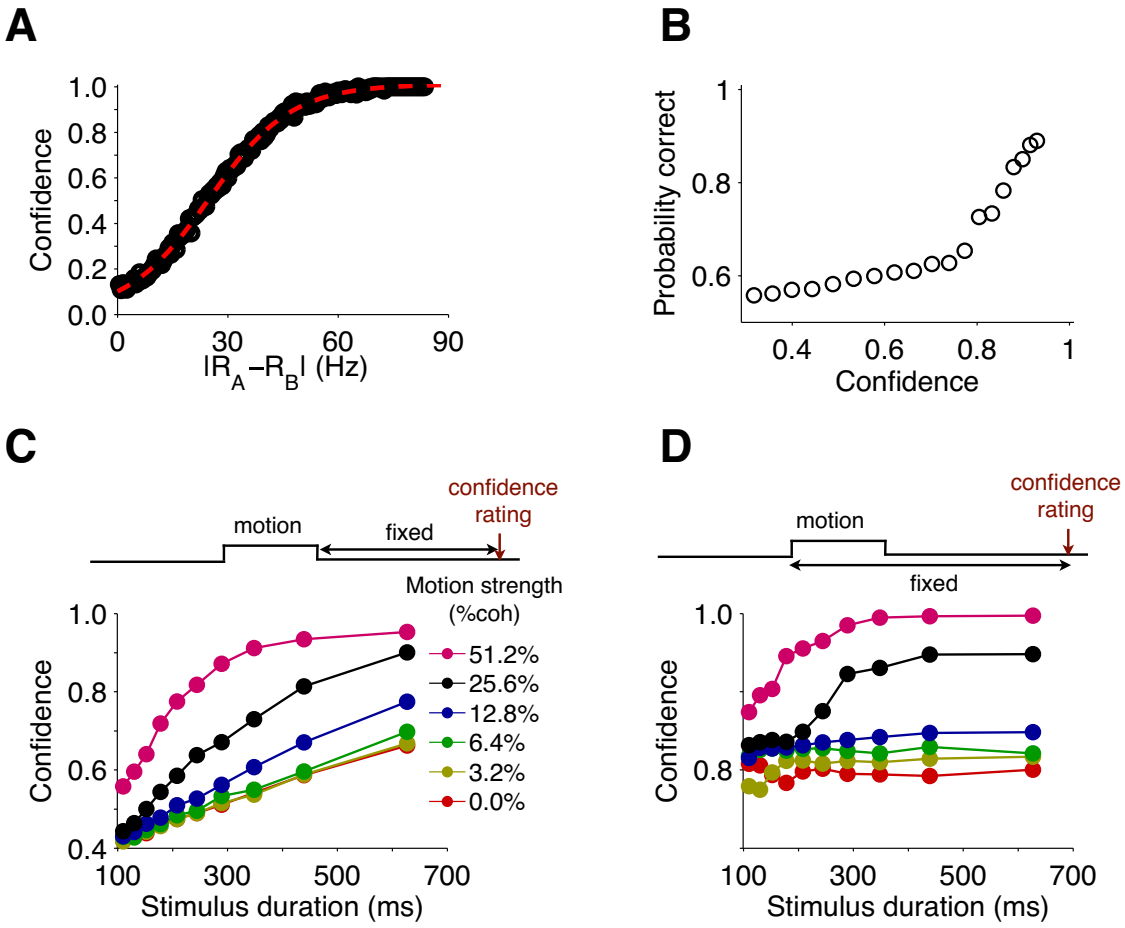


Figure 4.6: The probability of waiving T_s reflects choice confidence (**A**) Confidence is defined as the probability of waiving T_s at each $|R_A - R_B|$ level in single trials. A logistic function fit (red dash line) is performed on the data from **Figure 4.3B**. (**B**) Comparison of probability of correct and confidence at each $|R_A - R_B|$ level in single trials. Both confidence and probability of correct at each $|R_A - R_B|$ level in single trials are computed at decision time of trials without T_s presentation. Probability of correct increases as a monotonic function of confidence, which implies that confidence in our model would also be a good measurement of the subjective correct rate or log odds of choice (Kepecs and Mainen, 2012; Kepecs et al., 2008; Beck et al., 2008; Kiani and Shadlen, 2009; Drugowitsch et al., 2014; Drugowitsch et al., 2012). (**C**) The confidence assessment at T_s onset (the duration from motion offset to the time of confidence estimation is fixed, i.e. 575 ms, upper panel) increases with the motion strength and stimulus duration. (**D**) The confidence assessment at an identical time after the motion onset (the duration from motion onset to confidence estimation is fixed, i.e. 1550 ms, upper panel) saturates after a short period of stimulus duration. In this case, early evidence plays a dominant role in confidence estimation.

larger $|R_A - R_B|$ is, the closer is the system to a choice attractor T_A or T_B; $f(\cdot)$ is therefore required to be an increasing function. In the previous studies (Vickers, 1979; Kepecs and Mainen, 2012; Kepecs et al., 2008; Beck et al., 2008; Kiani and Shadlen, 2009; Drugowitsch et al., 2014; Drugowitsch et al., 2012), functions $f(\cdot)$ were given in a variety of ways. One can picture that as long as $f(\cdot)$ is a monotonic function, we can always equate $f(\cdot)$ from one model to another. Of note, our definition of the confidence stems from the structure of the attractor basin in decision space (**Figures 4.2G-H**), i.e. if a choice is confident, then it is more strongly resistant to the other external inputs such as a sure target, while confidences from models like Beck et al. (2008) and Kepecs and Mainen (2012) are compared directly to log odds of a choice in Bayesian framework. In the studies of Beck et al. (2008), they found both from the experimental data and their model that log odds of choice at A, namely confidence across trials for choice A, is proportion to $< R_A > - < R_B >$, where $< \cdot >$ is the average across trials. However, such a read-out of confidence would predict a strong correlation between confidence and performance on single trials, which is somewhat inconsistent with experimental observation of a broad performance variation in different confidence categories (Juslin and Olsson, 1996; Juslin and Olsson, 1997; Graziano et al., 2015; Graziano et al., 2010; Graziano and Sigman, 2009). Second, this "optimal decoder" $< R_A > - < R_B >$ relies explicitly on the equal variance hypothesis for likelihood (Kepecs and Mainen, 2012) or "left-right" symmetry of the linear decoder (Beck et al., 2008). It remains unclear the biologically plausible mechanism to achieve such a

fine-tuned neural circuit to compute confidence signal in these models.

In our model, confidence is defined using a monotonic increasing function of $|R_A - R_B|$. Particularly in this "opt-out" task, confidence can be probed by the probability that a choice stays in its attractor after presenting a sure target. If the choice is confident at $|R_A - R_B|$, then this probability, 1-P_{sure}, is low. Using this probe, we found that choice confidence increases as a sigmoid function (i.e. function $f(\cdot)$) of $|R_A - R_B|$ (**Figure 4.6A**). Next, we asked whether, across trials, our definition of confidence can also reflect probability of a correct choice at each $|R_A - R_B|$ level as those defined in Bayesian framework (Kepecs and Mainen, 2012; Kepecs et al., 2008; Beck et al., 2008). This seems possible as indicated from **Figure 4.5D**. Moreover, a detailed analysis should compare probability of correct choice and confidence simultaneously. We performed this analysis using trials without presenting T_s at decision time (**Figure 4.6B**). **Figure 4.6B** demonstrates that confidence in our model increases monotonically with the performance. Importantly, our model indicates that confidence can be computed as a function of the instantaneous neural activities (like a population code; (Beck et al., 2008)) at any time in a decision circuit without explicit use of elapsed time for integration of the sample (Moreno-Bote, 2010; Kepecs and Mainen, 2012; Kepecs et al., 2008; Beck et al., 2008; Kiani and Shadlen, 2009; Drugowitsch et al., 2014; Drugowitsch et al., 2012). Therefore, even though confidence in our model is not defined as a log odds function of the choice (Kepecs and Mainen, 2012; Kepecs et al., 2008; Beck et al., 2008; Kiani and Shadlen, 2009;

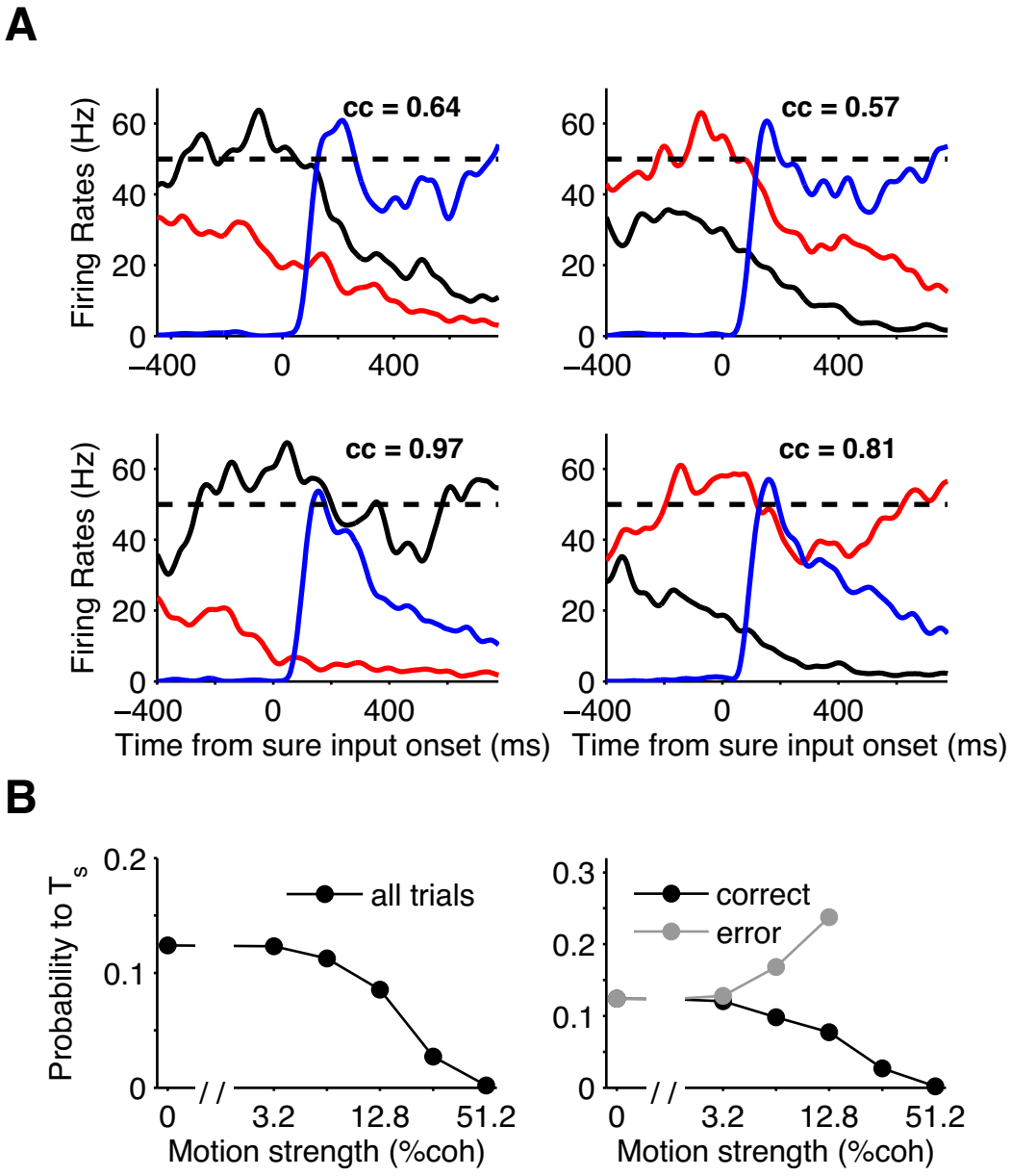
Drugowitsch et al., 2014; Drugowitsch et al., 2012), confidence can be a good measurement of the subjective correct rate across trials. Importantly, on single trials, choice confidence in our model is dissociable from performance (Graziano et al., 2015; Graziano et al., 2010; Graziano and Sigman, 2009), while Bayesian models would predict a strong correlation.

Despite the similarity of $f(\cdot)$, choice confidence in our model is however conceptually distinct from those from Bayesian decision models, since our definition of confidence fundamentally comes from the structure of the attractor basin in the decision space. Therefore, our model predicts that confidence would be different when estimated at the different times after motion offsets (**Figure 4.5**), whereas it would be nearly the same in a Bayesian model (Moreno-Bote, 2010; Kepecs and Mainen, 2012; Kepecs et al., 2008; Beck et al., 2008; Kiani and Shadlen, 2009; Drugowitsch et al., 2014; Drugowitsch et al., 2012). To test this, we estimated choice confidence using neural activities R_A and R_B in trials without T_s presented at different times after motion offsets. We first estimated the choice confidence at 575 ms after motion offset (**Figure 4.6C**; comparing directly to **Figure 4.3B**, left panel), where the distribution of $R_A - R_B$ is still evolving, namely the confidence estimation after a short delay (**Figure 4.5A**). One can thus expect an increase of choice confidence in trials with longer stimulus durations (**Figure 4.6C**), according to the variation of the bimodal distribution of $R_A - R_B$ at different stimulus durations (**Figures 4.4C-D**). We next estimated the choice confidence at 1550 ms after motion onset, i.e. the same time

of a trial, where the internal noise is nearly identical at different stimulus conditions, and the strength of the input dominates the network's choice confidence. In this case, one would expect that confidence should increase as a function of motion strength and stimulus duration for a noiseless integrator (Beck et al., 2008), unless it is bounded (Kepecs and Mainen, 2012; Kepecs et al., 2008; Kiani and Shadlen, 2009; Drugowitsch et al., 2014; Drugowitsch et al., 2012). While in our model, the attractor dynamics implies that the bimodal distribution of $R_A - R_B$ is dominated by the early divergence (**Figures 4.4C-D**). As a result, **Figure 4.6D** shows that all confidences saturate at stimulus duration > 400 ms, suggesting that the early evidence has the greatest effect on confidence estimation. Of note, the saturation time is longer with lower motion strength.

4.2.4 Low confidence results in changes of mind to sure target

The whole dynamic space, R_A-over-R_B decision space, can be divided into three regions: choice attractor regions ($R_A >> R_B$ or $R_A << R_B$) and an unstable region in-between them (Wang, 2008). In the previous study, we focused on the trials along the diagonal line ($R_A = R_B$), where a choice of network remains undecided before the presentation of T_s. We then investigated the dynamics of networks that are between the diagonal lines ($R_A = R_B$) and a choice attractor ($R_A >> R_B$ or $R_A << R_B$)


A
cc = 0.64
cc = 0.57
cc = 0.97
cc = 0.81
Firing Rates (Hz)
60
40
20
0
–400
0
400
Time from sure input onset (ms)
B
Probability to T_s
0.2
0.1
0
all trials
correct
error
0.3
0.2
0.1
0
0
3.2
12.8
51.2
Motion strength (%coh)

on R_A-over-R_B decision space preceding the presentation of T_s, where a trial could be still in the unstable region (and thus goes to the sure attractor after T_s onset) or in a stable attractor region (where the network trajectory stays in the same choice attractor even after T_s onset). For these trials, we could define an initial choice of the network by its nearby choice attractor T_A (T_B, respectively), where R_A (R_B, respectively) fires above a decision threshold (> 50 Hz). Particularly, we explored under which condition the network would more likely continue converging to the attractor of its initial choice, or shift to the sure attractor, when T_s is offered.

Figure 4.7A compares the neural activity in trials with low- and high-confidence initial choices. This analysis is performed on single trials and is missing in Kiani and Shadlen (2009). In low-confidence trials (**Figure 4.7A**, upper panels), one of the firing rates reaches a steady state and the other remains similar ($|R_A - R_B|$ is small). After T_s onset, both R_A and R_B decay to a low level, while R_s grows to a high level

Figure 4.7 *(preceding page)*: Low confidence results in changes of mind to sure target in post-decision wagering
(**A**) Trials with low-confidence exhibit changes of mind to T_s in PDW (motion strength: 6.4%; stimulus duration: 243 ms). Upper panels: sample trials with low-confidence, small $|R_A - R_B|$. Even though the network has reached one of the two choice attractors (left: T_A (black lines); right: T_B (red lines)), upon the presentation of T_s, the neural pool selective for T_s takes over (blue lines), so there are changes of mind. Lower panels: sample trials with high confidence, large $|R_A - R_B|$. No changes of mind take place. Choice confidence, cc, for each trial is estimated at the time of T_s onset and shown at the top of each panel. (**B**) Left panel: across trials (averaged over different stimulus durations), the probability of shifting to T_s decreases with the motion strength. Right panel: in error (correct) trials, this probability increases (decreases, respectively) with the motion strength. This prediction can be tested experimentally.

and T_s is chosen. By contrast, in high-confidence trials (**Figure 4.7A**, lower panels), one of the firing rates reaches a steady state and the other is much smaller ($|R_A - R_B|$ is large). Neurons activated by T_s are suppressed, and T_s is waived. In the latter case, the activity of the winning neural pool exhibits a brief dip upon T_s onset, and then ramps up again to its steady state.

Across trials, the probability of changes of mind to T_s is negatively correlated with choice confidence, i.e. the network exhibits low confidence in trials at low motion strength (**Figure 4.6B**), and high probability of changes of mind to T_s (**Figure 4.7B**, left panel). To further test whether the network bases the probability of changes of mind to T_s on its performance and confidence, we categorized the trials with initial choices, where either R_A or R_B reaches a decision threshold, 50 Hz (if both of them do not reach the decision threshold, we considered the choice remaining undecided at T_s onset), into correct and error groups, and found that network changes its choice to T_s more often in error trials. Furthermore, the probability of choosing T_s in correct (error, respectively) choice decreases (increases, respectively) with the motion strength (**Figure 4.7B**, right panel). This finding is reminiscent of the experimental observation that, in a PDW task with a delayed reward, animals moved back to self-restart port more often when they made an erroneous choice (Kepecs et al., 2008).

In conclusion, we identified two possibilities for choosing T_s: either an initial choice was not made (along the diagonal line; **Figure 4.2**), or it was made with low confidence (between the diagonal line and choice attractors; **Figure 4.7A**, upper

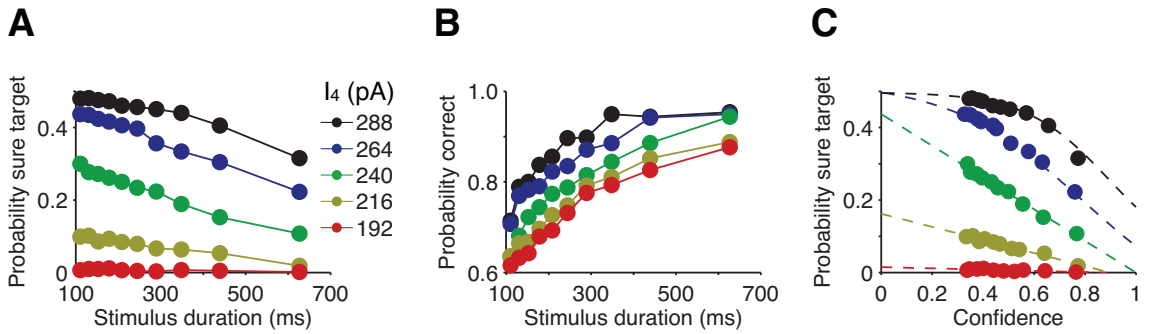


Figure 4.8: Effect of sure target input strength on the behavioral performance
In this simulation, the motion strength is fixed at coh=12.8%, and T_s input strength at $I_4 = 240$ pA (green circles and line) is the same as those used in Figures 2-7. (**A**) P_{sure} increases as a function of T_s input strength. T_s is usually waived (chosen, respectively), when T_s input strength is weak (strong, respectively). (**B**) Correct rate in the trials, where T_s is waived, increases as a function of T_s input strength. (**C**) Choice confidence is identical at the moment of T_s onset (but increases as a function of stimulus duration). For a range of T_s input strength (216 pA $< I_4 <$ 264 pA), P_{sure} decreases as a linear function of choice confidence and thereby can be considered as a probe about choice confidence.

panels). For the latter case, $|R_A - R_B|$ reveals the confidence about an initial choice; low confidence of a choice is likely to result in changes of mind to T_s.

4.2.5 A sure target as a probe about the system's confidence

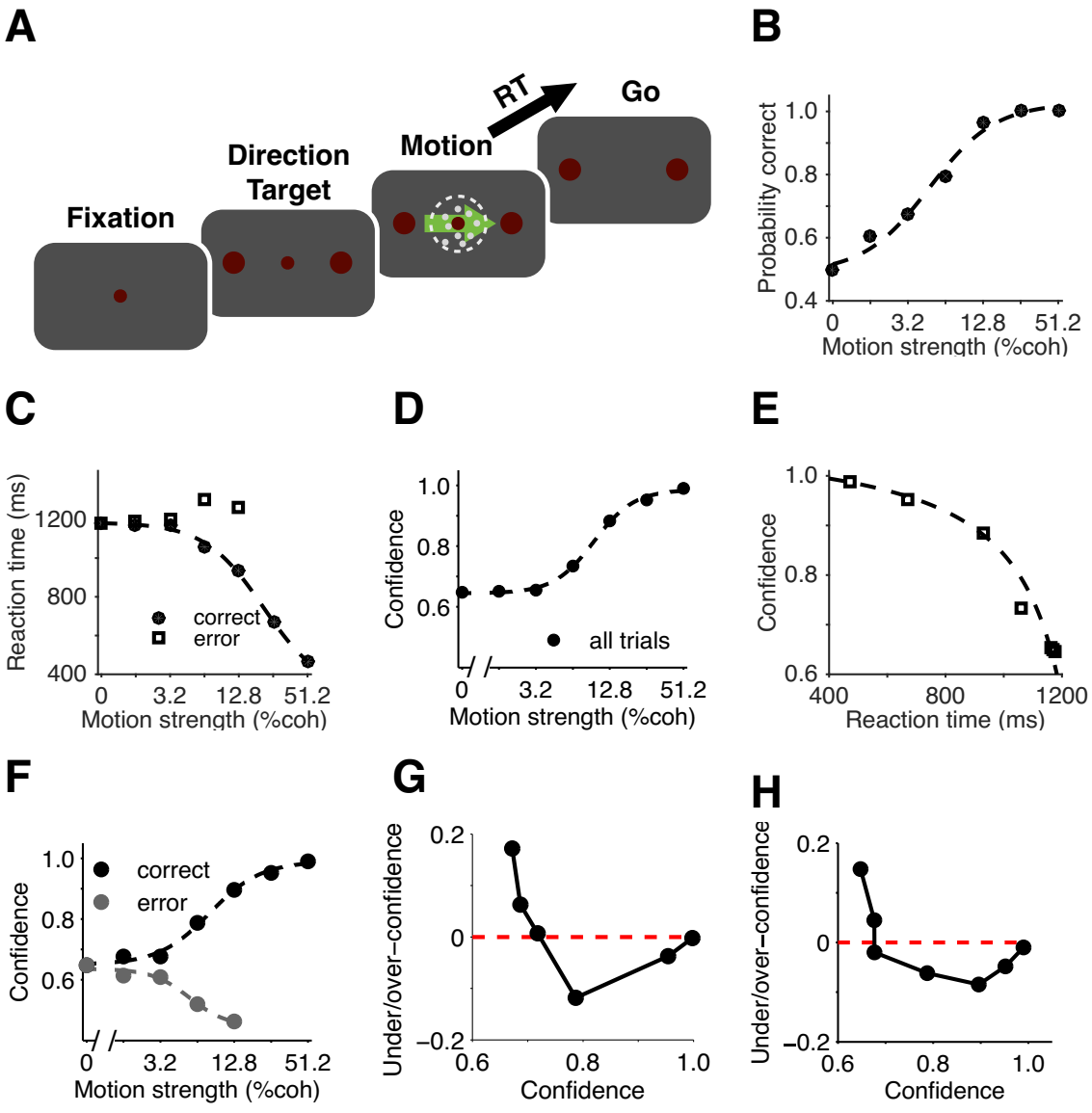
The introduction of a sure target plays a role of probing the system's confidence. Specifically, in the monkey experiment, the physical luminance of the sure target was the same as the choice targets. Monkeys were trained to understand what the sure target meant behaviorally, which depended on the amount of reward by choosing it.

Therefore, in our model, the amplitude of the sure target input (I_4 in our model) does not correspond to its physical properties but is related to the behavioral significance of the sure target that a monkey learned as the amount of reward he receives by choosing the sure target. One can imagine that if choosing the sure target yields a negligible (significant, respectively) amount of reward, monkeys would never (always, respectively) have learnt to choose it. To test this, we studied the effect of T_s input strength, I_4 (**Materials and Methods**, **Equation 4.6**), on the behavioral performance at a fixed motion strength level (i.e. 12.8%). We found that P_{sure} increases as a function of T_s input strength (**Figure 4.8A**). When T_s input strength is low (e.g. $I_4 = 192$ pA), T_s is always waived; when T_s input strength is high ($I_4 = 288$ pA), T_s is mostly chosen, as stimulus duration is short. Moreover, in the trials where T_s is shown but waived, our simulation predicts an increase of correct rate at high T_s input strength (**Figure 4.8B**). At the network level, these observations in **Figures 4.8A-B** still follow a three-way competition among R_A, R_B and R_s, e.g. when input of the sure target is weak (strong, respectively), it always behaves like a loser (winner, respectively). Last, we examined whether in a range of T_s input strengths (a selected range of amount of T_s rewards), a sure target can serve as a probe about the systems confidence, when the network applies the attractor dynamics. We assessed the choice confidence as a function of $|R_A - R_B|$ at the moment of T_s onset for different choices of T_s input strengths. **Figure 4.8C** shows that, on average, choice confidence is identical for different T_s input strengths (and increases as a function of stimulus

durations), and P_{sure} decreases as a linear function of choice confidence for a broad range of T_s input strengths, i.e. in this range, 216 pA $< I_4 <$ 264 pA, a sure target in our model can be considered as a probe about the system's confidence. Therefore, a sure target is only a probe and the confidence measure is valid even without it.

4.2.6 Assessment of choice confidence in a reaction time task

In our model, confidence can be read out at any time and increases as a function of stimulus duration in an FD task. One may thus argue that the network would exhibit high confidence despite the task difficulty if it freely controls the viewing duration of the stimulus. However, classical literature about confidence in Cognitive Psychology (Vickers, 1979) emphasizes an inverse relationship between confidence and response time, which can be potentially tested in a reaction time (RT) version of discrimination task (developed previously by Furman and Wang (2008)) with direct assessment of choice confidence. This distinguishable difference between confidence estimation in FD vs RT task in fact comes from two distinct processes, whereas longer viewing time in a FD task enables more integration of evidence (confidence thus increases with motion viewing time), a longer RT means a higher task difficulty in a RT task (confidence thus decreases instead with motion viewing time). We thus want to further test whether our model can nicely explain such a contrasting observation. To


A
Fixation
Direction Target
Motion
RT
Go
B
Probability correct
1.0
0.8
0.6
0.4
0
3.2
12.8
51.2
Motion strength (%coh)
C
Reaction time (ms)
1200
800
400
correct
error
0
3.2
12.8
51.2
Motion strength (%coh)
D
Confidence
1.0
0.8
0.6
all trials
0
3.2
12.8
51.2
Motion strength (%coh)
E
Confidence
1.0
0.8
0.6
400
800
1200
Reaction time (ms)
F
Confidence
1.0
0.8
0.6
correct
error
0
3.2
12.8
51.2
Motion strength (%coh)
G
Under/over−confidence
0.2
0
−0.2
0.6
0.8
1.0
Confidence
H
Under/over−confidence
0.2
0
−0.2
0.6
0.8
1.0
Confidence

do this, we designed a reaction time (RT) version of discrimination task with direct assessment of choice confidence (**Figure 4.9A**): the network integrates the motion input until the neurons selective for one of two alternatives fire above a decision threshold, and reports the confidence as the function of the instantaneous $|R_A - R_B|$ (**Figure 4.6A**) at the moment of choice (a similar human behavioral experiment is performed and reported by Kiani et al. (2014) recently).

Our model exhibits the typical psychometric and chronometric curves of a two-alternative discrimination task (Roitman and Shadlen, 2002; Churchland et al., 2008), i.e. $P_{correct}$ increases, while RT decreases with the motion strength (**Figures 4.9B-C**). Importantly, weaker motion strengths are associated with longer RT_s, where $|R_A - R_B|$ will be less at longer RT_s. Choice confidence thus increases with the motion

Figure 4.9 ***(preceding page)*****: Choice confidence in a reaction time task**
(**A**) Reaction-time (RT) discrimination task with confidence rating. In task, a subject can indicate its choice at any time after the motion onset simultaneously with a direct report of confidence. (**B**) Psychometric and (**C**) chronometric curves. $P_{correct}$ increases while RT decreases with the motion strength. (**D-F**) Confidence reported as a post-hoc feature of decision. (**D**) Choice confidence increases with motion strength (see also the result in Figure 5 (Beck et al., 2008)). (**E**) Confidence decreases as an inverse function of RT ($cc = \frac{a}{t-b} + c$; a=91.24 ms, b=1369.35 ms, c=1.089 are parameters to fit, $R^2 = 0.998$; black line). (**F**) Confidence increases (decreases) as a function of the motion strength in correct (error, respectively) trials. Figures **B** and **D** imply that choice confidence increases with $P_{correct}$. We found that for the low accuracy case, the simulation exhibits overconfidence (confidence estimation is greater than correct rate), while for the high accuracy case, it exhibits underconfidence (confidence estimation is lower than correct rate). (**G-H**) Variation of under/over-confidence score with the increase of confidence in the FD task with a delay of 627 ms and RT task, respectively. The network behaves with overconfidence (above zero) in very difficult trials (at zero, 3.2% and 6.4% motion strengths for FD task; at zero and 1.6% motion strengths for RT task), but with underconfidence (below zero) in easy and moderately difficult trials.

strength (**Figure 4.9D**; see also Figure 5 in (Beck et al., 2008)) and is positively correlated with the behavioral performance across trials (Barthelmé and Mamassian, 2010) (data not explicitly shown). We also found that choice confidence decreases as an inverse function of RT (**Figure 4.9E**), which agrees broadly with the human behavioral observations (Vickers, 1979). Even though an erroneous choice could be associated with high confidence (Graziano and Sigman, 2009), the average $|R_A - R_B|$ across trials is higher in correct trials than that in error ones (Wang, 2002). Therefore, in our model, confidence increases (decreases, respectively) with motion strength in correct (error, respectively) trials (**Figure 4.9F**), consistent with human studies (Pierrel and Murray, 1963).

Moreover, we studied correlation between the choice confidence and decision accuracy. **Figure 4.9B** and **Figure 4.9D** imply that choice confidence is positively correlated with behavioral performance across trials. Although confidence in our model does not directly represent a subjective estimation of performance (like that in (Beck et al., 2008; Kiani and Shadlen, 2009; Drugowitsch et al., 2012)), one can estimate the subjective performance from choice confidence using the monotonic function, $g(\cdot)$, in **Figure 4.6B**. We can thus compare directly our confidence score with performance to study the "hard-easy" effect (Juslin and Olsson, 1997). Here we defined underconfidence score as the difference between the choice confidence and accuracy, $cc - P_{correct}$ (one can also use $g(cc) - P_{correct}$), and a "hard-easy" effect is the observation that the underconfidence score decreases as a function of task diffi-

culties, i.e. in the easy (difficult, respectively) trials, the report is more likely to be overconfident (underconfident, respectively) $cc - P_{correct} > 0$ ($cc - P_{correct} < 0$, respectively). **Figures 4.9G-H** show the variation of underconfidence scores as a function of choice confidence for the FD task with a delay of 627 ms and RT task, respectively; both display the "hard-easy" effect in the reports. Of note, these results still hold true when comparing estimated subjective performance from choice confidence in our model with the behavioral performance, i.e. $g(cc) - P_{correct}$. Such a "hard-easy" effect in our model mainly stems from sampling of stochastic neural dynamics; sampling duration thus influences the underconfidence score in our model. When the sampling duration is short, the network behaves with more overconfidence. To test this, we compared the scores in the FD and RT tasks, where the average sampling durations in RT tasks are longer than those in FD tasks at low motion strength (from zero to 6.4%). Consequently, the network exhibits overconfidence more often at low motion strengths in FD task.

4.3 Discussion

We have shown that a biologically plausible spiking network model can account for salient physiological and behavioral data from an experiment designed to study confidence (Kiani and Shadlen, 2009), and in our model internal uncertainty plays an essential role of choice confidence (see also (Whiteley and Sahani, 2008)). Specifically,

at the moment of choice, our model simultaneously generates a neural signal for confidence. Confidence can be estimated as a function of the differential activity of the competing neural populations, $|R_A - R_B|$. Comparing to Bayesian inference models, in our model, there is no explicit representation of probabilities such as likelihood or posterior function. Indeed, all computations are carried out by the fluctuating neural network dynamics. Therefore, confidence estimation itself is simply a quantity that stochastically varies over time, and from trial to trial under the same stimulus condition.

Our identification of a confidence signal, $|R_A - R_B|$ agrees with the idea that as a metacognitive process, confidence is estimated directly during a decision process (Graziano et al., 2015; Graziano et al., 2010; Graziano and Sigman, 2009; Middlebrooks and Sommer, 2011; Middlebrooks and Sommer, 2012). At the same time, choice confidence is also dissociable from whether the decision is correct or wrong in a single trial, as illustrated by high-confidence error trials (**Figures 4.7A**, lower right panel). In line with our model, the EEG data from (Graziano et al., 2010) showed that at the neural level, choice confidence could be dissociated from performance. Such dissociation is naturally explained by attractor dynamics, which could yield the same magnitude of the differential activity $|R_A - R_B|$, hence the same confidence rating in correct and error trials. It is worth noting that (1) R_A and R_B represent the choices of a decision (not necessarily a directional decision making process); (2) confidence estimation does not depend on a specific choice of the decision (i.e. not

exclusively rely upon the activity of the winner bump, nor the losing bump), but a differential activity between choices. In this case, one would expect a sure target is chosen when $|R_A - R_B|$ is small or the downstream neuronal activity is weak, and a non-sure target is chosen when $|R_A - R_B|$ is large or the downstream neuronal activity is strong. This prediction from our model is consistent with the observations in Komura et al. (2013), wherefore the finding of pulvinar neuronal activity (Komura et al., 2013) could be an example of $|R_A - R_B|$ in the downstream read-out circuit of confidence.

In our model, fast early divergence, i.e. the difference of early buildup rates between R_A and R_B, has the predominant effect on the choice and confidence. This is manifested in the dependence of the choice confidence on the stimulus duration, which saturates quickly for sufficiently long stimuli (**Figures 4.6D**) (Wong et al., 2007). By contrast, in DDM, sensory evidence contributes equally in time to confidence estimation. Future experiments are needed to assess this different characteristic of the attractor network model versus DDM. Furthermore, the two competing neural pools could also diverge slowly later in a trial. In our model, persistent activity during the delay not only maintains working memory, but also continues to slowly integrate sensory signals from memory (Curtis and Lee, 2010). This provides a neural mechanism for post-decision sampling (Resulaj et al., 2009). For instance, **Figures 4.7B** shows that the probability of switching from an initial decision to T_s is higher in error trials, in agreement with behavioral observation in a rat experiment (Kepecs

et al., 2008). This finding also sheds insights into the phenomenon of changes of mind, which may result from the instability (low confidence) of a choice (see also (Albantakis and Deco, 2011)).

Of note, in the monkey experiment, as well as in our model simulations, the introduction of a sure target only serves as a probe about the systems confidence (**Figures 4.8**). The probability of opting for the sure target is bounded (**Figures 4.3** and **Figures 4.8C**), so it represents a good choice for estimating confidence. The real result, we emphasize here, is to quantify confidence as a function of the neural activity $|R_A - R_B|$. Confidence thus quantified should be applicable to all trials, even without sure target presentation. Furthermore, in Kiani-Shadlen's analysis, they also found that the probability of opting for the sure target can be predicted using either $f(|R_A - < R_A >|)$ or $f(|R_B - < R_B >|)$. Nevertheless, $f(|R_A - < R_A >|)$ or $f(|R_B - < R_B >|)$ is not a good measure of confidence for a reaction time task, for which either R_A or R_B is assumed to reach a fixed threshold at the moment of the choice, therefore $f(|R_A - < R_A >|)$ or $f(|R_B - < R_B >|)$ would always be a fixed value ($f(|\text{threshold} - \text{average}|)$) rather than a graded quantity that varies from trial to trial (Kiani et al., 2014).

4.3.1 Comparison with existing models

Computational schemes have been proposed for the study of confidence (Vickers, 1979; Kepecs et al., 2008; Ratcliff and Starns, 2009; Kiani and Shadlen, 2009; Rolls et

al., 2010a; Rolls et al., 2010b; Moreno-Bote, 2010; Kepecs and Mainen, 2012). These models can be classified into Bayesian inference models and neural network models.

In Bayesian inference models, one can either compute confidence based on a single decision variable (Kiani and Shadlen, 2009; Drugowitsch et al., 2012) or the optimal population code (Beck et al., 2008). Kiani and Shadlen (2009) proposed that confidence could be defined in terms of the log posterior ratio for the two choices given the position of a decision variable and elapsed time at decision time, using DDM. This looks promising, yet it remains unclear what is a direct representation of a decision variable exclusively for a choice. Moreover, for RT version of the task, this kind of models implies the position of a decision variable at decision time would be a deterministic function of RT (either a constant or a time-varying function like that in (Drugowitsch et al., 2012)); one can thus find that confidence would also decrease deterministically as a monotonic function of RT (Volkmann, 1934; Kiani and Shadlen, 2009; Drugowitsch et al., 2012) on single trials. This idea, however, failed to explain the widely overlapped RT distributions in different confidence categories (Ratcliff and Starns, 2009). Such a strong correlation between confidence and RT or performance can be eliminated through a two-stage DDM (Pleskac and Busemeyer, 2010), where additional process for confidence is required. Nevertheless in our model, performance, RT and confidence are naturally dissociated with each other on single trials. Importantly, in a classic DDM model, sensory evidence contributes equally in time to confidence estimation, while in our model, confidence estimation is more dominated

by the early sensory evidence. Last, in Kiani and Shadlen model, choosing T_s is a hard thresholding process and has little to do with neural activity at T_s response field, while in our model, it is generated from the same sampling of stochastic neural dynamics as the other choices (as indicated by data in Figure 5, (Kiani and Shadlen, 2009)).

On the other hand, the optimal population code model (Beck et al., 2008) claimed that confidence could be estimated as the instantaneous differential activity $|R_A - R_B|$, without explicit use of RT as our model. A notable difference between our model and theirs is that the optimal population code model requires the LIP neural circuit to be a fine-tuned noiseless integrator. This can be easily tested experimentally, since our model predicts that confidence estimation would differ at different times in the delay, while their model would expect it to be constant. Generally, these Bayesian inference models (Beck et al., 2008; Kiani and Shadlen, 2009) claimed that confidence must be based on explicit neural representation of probability functions such as likelihood at any moment in time and in single trials. Our model demonstrates, convincingly, that this might be an inaccurate perspective. Whereas probability representations may be a perfectly valid mathematical description of the aggregated statistics across trials, they should not be confused with what actually happens in single trials, which is stochastic neural dynamics.

Furthermore, Deco and his collaborators have recently developed a variety of spiking neuron based network model to account for the confidence estimation and

its behavioral readout (Insabato et al., 2010; Rolls et al., 2010a; Rolls et al., 2010b). Insabato et al. (2010) argued that confidence can be read out as a function of $R_A + R_B$ and Rolls et al. (2010a); Rolls et al. (2010b) claimed that it can further approximated as a function of the neural activity of the winning pool. All these models showed some consistencies with the existing data. However, as discussed in our model, neither of them can exclusively demonstrate the position of a neural trajectory related to the choice attractor and thus the choice confidence at any time during a decision. These results seem only true at moment when a decision is made exactly around a choice attractor in the decision space, where $|R_A - R_B| \simeq |R_A + R_B| \simeq \max(R_A, R_B)$, since $\min(R_A, R_B) << \max(R_A, R_B)$. Therefore, these models would fail to predict confidence using a fixed decision boundary, nor can they capture the relationship between neural activities in LIP with P_{sure}, or high-confidence errors in single trials. Alternatively, our model does not require a time-varying decision threshold, estimates confidence simply as a function of instantaneous $|R_A - R_B|$ at the moment of choice on single trials, and can correctly reproduce the salient behavioral relationships between confidence, RT and performance on single trials and those across trials.

Confidence rating is important for monitoring cognition when there is uncertainty, and two types of uncertainty should be distinguished: Brunswikian (external) uncertainty originating from incomplete states of knowledge (noisy or ambiguous sensory data), and Thurstonian (internal) uncertainty due to variations intrinsic to the brain (Juslin and Olsson, 1997). The noise level in a decision circuit has only recently

begun to be examined experimentally (Brunton et al., 2013). Our work provides a computational framework to detects these two effects using a spiking-neuron circuit. Our model can also be extended in several important ways. It still remains an open question how confidence estimation, as a sigmoid function of the differential activity in downstream neural circuits, can be read out for a direct report and to guide future behavior. In fact, confidence is commonly assessed without a verbal report using a two-stage PDW task: subjects perform a first-order discrimination task, and then make a high-low bet upon the outcome of the decision (Smith, 2009; Middlebrooks and Sommer, 2011; Middlebrooks and Sommer, 2012) (see also Kiani et al., 2011, Soc. Neurosci. Abstr., 17.06), where the probability for a high bet is considered as a readout of confidence estimation. A plausible neural circuit for explicit representation and memory of a confidence signal is needed for the two-stage PDW (Smith, 2009; Middlebrooks and Sommer, 2011; Middlebrooks and Sommer, 2012; Komura et al., 2013) and should be examined in the future. A biologically plausible neural circuit to computing $|R_A - R_B|$ involves neurons in pulvinar (Komura et al., 2013), where the neurons fires at high rates in non-sure target trials, and at low rates in sure target trials. Moreover, Kiani et al. (2014) recently found that confidence could also decrease with the difficulty of task, which poses a challenge to our model prediction (**Figures 4.9F**). One possible direction in the future is to understand the mechanism of error trials. Nevertheless, a robust prediction of our model, compared to their observations, is that the difference of the confidence between correct and

error decreases with the difficulty of task (Lak et al., 2014). Secondly, for the sake of simplicity, we assumed the amount of T_s reward is encoded as the onset strength of its target input, which mimics the firing activity of midbrain dopamine neurons in response to the targets with different amounts of reward (Tobler et al., 2005). Our model predicts that both P_{sure} and $P_{correct}$ increase with the reward of T_s (data not shown). This then brings up two questions in future studies: (1) what would be a reasonable amount of T_s reward used to measure confidence in a PDW task (Persaud et al., 2007; Dienes and Seth, 2010; Fleming and Dolan, 2010), and (2) how the amount of T_s reward obtained is learnt through neural dynamics and applied to the decision circuit (Soltani and Wang, 2006). Moreover, one can extend our model to investigate confidence signals for multiple-choice decision tasks and effects of microstimulation on confidence (Fetsch et al., 2014). Specifically, one can incorporate known effects of micro-stimulation on MT inputs in our model to perform the experiment of Fetsch et al. (2014) using computer simulation and then test its effects on confidence. Finally, confidence may be represented in a distributed network in the brain (Del Cul et al., 2009), the dynamical nature and computational principle remains to be elucidated in future research. In conclusion, we found it remarkable that a previously established model of decision-making (Furman and Wang, 2008) naturally accounts for all the salient behavioral and neurophysiological observations of the Kiani-Shadlen experiment. Furthermore, it reproduces the observation that confidence decreases with response time in a reaction time version of the task. The model also offers testable

predictions about the changes of mind and, unexpectedly, the "hard-easy effect" observed in human studies, which naturally emerges from the model. Taken together, our work establishes that a dynamical system of stochastic neural population can underlie even the seemingly abstract metacognitive concept of confidence.

4.4 Materials and Methods

4.4.1 Network model

We employed a spiking neural network model, which has been previously used to simulate a categorical decision of an analog feature, like motion direction (Furman and Wang, 2008; Liu and Wang, 2008). This model consists of 2048 pyramidal cells and 512 interneurons. Both pyramidal cells and interneurons are modeled as integrate-and-fire neurons; excitatory postsynaptic currents from pyramidal cells are mediated by models of AMPA and NMDA receptors, while inhibitory postsynaptic currents from interneurons are mediated by GABA receptors. Pyramidal cells are uniformly placed on a ring according to their preferred motion directions and continuously span 360 degrees of possible motion directions (**Figure 4.1B**), while the interneurons constitute a non-selective neural pool. The recurrent connectivity strength between two pyramidal cells is a Gaussian function of the difference between their preferred motion directions, while those from and onto the interneurons are broad and uniform (**Figure 4.1B**). Each cell receives an independent background noise mediated by

AMPA receptors, which is modeled as uncorrelated Poisson spike trains. We used the neuronal and synaptic parameters from (Furman and Wang, 2008), which are fully specified therein, with a change of the background noise rate to 2200 Hz, which ensures a choice is generated even if the motion strength is weak and stimulus duration is short (this mimics behavioral results in Kiani-Shalden's experiment). With these parameters, the network is endowed with winner-take-all competition so that only one of the neural pools wins (reaching an average population firing rate > 50 Hz for at least 50 ms), and the decision is maintained in the form of a bell-shaped persistent activity pattern ("bump attractor") during the delay period.

4.4.2 Simulation protocol of fixed-duration discrimination decision task

Our model assumes that neurons in area LIP incorporate sensory evidence (Cook and Maunsell, 2002; Roitman and Shadlen, 2002; Hanks et al., 2006) and reward signals (Platt and Glimcher, 1999; Sugrue et al., 2004; Tobler et al., 2009). For simplicity, we assumed that the amount of reward for each target (i.e. two directional targets and a sure target) is associated with the instantaneous input strength of its current at the moment of the target onset (Soltani and Wang, 2006). That is to say, the amplitude of the sure target input does not correspond to its physical properties (like the luminance) in the experiment, instead it is related to the behavioral signifi-

cance of the sure target that a monkey learned, i.e. the amount of reward it received by choosing the sure target (**Figure 4.8**).

In a fixed-duration (FD) version of the two-alternative direction discrimination task (**Figure 4.1A**), two directional targets, T_A (90°) and T_B (270°), are first presented to the network. A random-dot stimulus with net motion to T_A is presented at 500 ms after the two targets onset. The difficulty of the task is modulated by the stimulus duration (randomly chosen from 110 ms, 130 ms, 152 ms, 178 ms, 208 ms, 244 ms, 289 ms, 348 ms, 439 ms, and 627 ms), and the percentage of coherently moving dots (the motion strength). We modeled the external input to pyramidal cell i (at θ_i) as a sum of two target signals, $I^i_{tar}(t)$ (tar=$\{A, B\}$; **Figure 4.1D**, black line), and the motion stimulus, $I^i_m(t)$ (**Figure 4.1C**). The target inputs to T_A and T_B are identical:

$$I^i_{tar}(t) = I_{tar}(t) \exp[-\frac{(\theta_i - \theta_{tar})^2}{2\theta^2_{tar}}] \tag{4.2}$$

where $\theta_A = 90°$; $\theta_B = 270°$; $\theta_{tar} = 10°$.

$$I_{tar}(t) = \begin{cases} I_1 + I_2 \exp[-\frac{t-t_d-200}{\tau_d}], & \text{if } t_d + 200 < t < t_m + 80 \\ I_3 + (I_1 - I_3) \exp[-\frac{t-t_m-80}{\tau s}], & \text{if } t \geq t_m + 80 \end{cases} \tag{4.3}$$

where t_d = 400 ms and t_m = 800 ms are the onset times of targets and motion, respectively; τ_d = 500 ms and τ_s = 15 ms are the time constants of the adaption and suppression, respectively; I_1 = 250 pA, I_2 = 50 pA, and I_3 = 60 pA. Specifically,

the target (motion, respectively) input onset time is 100 (200, respectively) ms after the target (motion, respectively) onset time, and the target input is suppressed by the motion stimulus with a latency of 80 ms (Roitman and Shadlen, 2002); with the high intensities of the target inputs, so winner-take-all competition between the two targets does not take place prior to the motion stimulus onset (Wong et al., 2007; Furman and Wang, 2008; Liu and Wang, 2008).

In simulation, motion input is modeled to imitate the neural response in the middle-temporal area to the random-dot stimuli. We constructed such a population activity as a Gaussian function with a tuning width independent of motion strength while motion presented ($t_m + 200 < t < t_{mo}$, t_{mo} is the moment of motion input offset)

$$I_m(i) = m_0 + coh\{-m_1 + m_2 \exp[-\frac{(\theta_i - \theta_m)^2}{2\sigma_m^2}]\} \tag{4.4}$$

where the motion strength $0 \leq coh \leq 1$; net direction $\theta_m = 90°$; $\sigma_m = 40°$. We kept the activity normalized, i.e. $< I_m(i) >= m_0 = 4$ pA; $m_1 = 4.93$ pA; $m_2 = 25$ pA.

In trials with T_s ($\theta_s = 180°$), where there was the opt-out safe target presented, (**Figure 4.1D**, red line), we modeled its time-dependent current, $I_s(i)$, as:

$$I_s(i) = I_s(t) \exp[-\frac{(\theta_i - \theta_s)^2}{\sigma_{tar}^2}] \tag{4.5}$$

which is added to the external input. We used:

$$I_s(t) = I_3 + I_4 \exp[-\frac{t - t_s - tmo}{\tau_{ss}}] \tag{4.6}$$

for $t > t_s + t_{mo}$, where t_s is T_s input onset time after the motion input offset, t_{mo}, with a latency of 100 ms to the network after T_s onset. In simulations, we used t_s=575 ms; τss = 90 ms; I_4=240 pA (see **Figure 4.8** for a discussion on the choice of I_4), expect **Figure 4.5** (t_s is equal to 575 ms, 750 ms, or 925 ms).

The network model is taken from Furman and Wang (2008), with a few parameter changes, i.e. background noise that ensures a choice is generated even if the motion strength is weak and stimulus duration is short, and a different set of parameter values for the choice target input, motion input and sure target input that are adopted to the new experimental protocol of Kiani-Shadlen experiment. Although the network was not originally designed for confidence estimation experiment, unexpectedly it can reproduce the behavioral and neurophysiological observations that are similar to those in Kiani-Shadlen experiment (Figure 1B-C; Figure 2A-B; Figure 5B-C in Kiani-Shadlen paper): (1) neurons inside T_A and T_B display indistinguishable firing activity when T_s is chosen, while their firing activity are divergent when T_s is shown but waived (**Figures 4.2B, D**, upper panel, versus Figure 2A-B in (Kiani and Shadlen, 2009); (2) neurons inside T_s response field have weak and uninformative spontaneous activity before T_s onset, and then exhibit a fast ramping followed by a decay after T_s onset

(**Figure 4.2D**, lower panel, versus Figure 5B-C in (Kiani and Shadlen, 2009)). In the model, we assumed that the input strength of each target goes to the same level due to the adaptation. For simplicity, we fixed I_4, while adjusting s to match a majority of points on the performance curves of P_{sure} from (Kiani and Shadlen, 2009) (**Figure 4.3B**, versus Figure 1B-C in (Kiani and Shadlen, 2009)).

We also studied the choice confidence in a reaction time (RT) version of the task. In this task, a network can generate a choice at any time after the motion onset, and at the same time, report directly its choice confidence. We followed the same simulation protocols as those in the fixed duration task without T_s, except that the motion input is terminated when one of the activity bumps crosses the decision threshold, 60 Hz for at least 50 ms. We measured the corresponding time, t_r, and calculated RT as $t_R = t_r - t_m + 80$, where 80 ms is the latency period for implementation of saccadic eye movement (Roitman and Shadlen, 2002). In the fixed-duration (FD) version of the task, an initial choice is assumed to be made when one of the two competing neural pools reaches a decision threshold of 50 Hz for at least 50 ms after motion onset, since the decision threshold of an FD task was experimentally observed lower than that of a RT task (Roitman and Shadlen, 2002).

In FD tasks, we performed 1500 trials at each motion strength and stimulus duration level, where T_s was not presented, and 3500 trials at each motion strength and stimulus duration level, where T_s was presented. In **Figure 4.5**, we simulated 1500 trials for each data point. In RT task, we carried out 3000 trials for each motion

strength level. All the simulated behavioral data reported were computed using all trials for each simulation set. The integration method was a modified second-order Runge-Kutta algorithm with firing-time interpolation (Hansel et al., 1998), and a time step $dt = 0.02$ ms.

4.4.3 Measurements of activity trajectories

We calculated the average response of the population of units associated with targets T_A, T_B, and T_s, namely R_A, R_B, and R_s, as the average firing activity of the neurons within 8.4° around each target with a time window of 100 ms preceding the time point (e.g. the moment of decision and onset of T_s) for R_A, R_B, and R_s in analysis, except for **Figures 4.2B, 4.2D-H, and 4.7A**. In **Figures 4.2B, 4.2D-H, and 4.7A**, each trajectory or point represents the activity of a single neuron at each target. We applied a 100 ms Gaussian sliding window to smooth the PSTHs for the temporal evolutions of the firing rates of R_A, R_B, and R_s in **Figures 4.2B, 4.2D-H, and 4.7A**.

4.4.4 Choice confidence assessment

In the monkey experiment, as well as in our model simulations, the introduction of a sure target only serves as a probe examining the system's confidence. That is to say, with carefully choosing the ratio of sure target reward to that of choice targets

(i.e. I_4 in our simulation), one can access the choice confidence across trials. In our simulation, the probability of opting for the sure target is bounded (**Figure 4.3B**); it thus represents a good choice for estimating confidence. Furthermore, we will show in the **Results** section that probability of choosing the sure target, P_{sure}, reflects the uncertainty of a choice in an opt-out task. Here, we defined choice confidence, cc, as the probability of waiving a sure target, i.e. $cc = 1 - P_{sure}$ at each differential activity level, using the trials in which T_s is presented (binned by 0.5 Hz; **Figure 4.6A**, black circles). We also assumed that this probability can be predicted as a function of the differential activity $|R_A - R_B|$. We then performed the fit of a sigmoid function between $|R_A - R_B|$ and $cc = 1 - P_{sure}$. At each differential activity level, i, $cc^i = 1 - P_{sure}$ is computed as the mean of the decision result across the sample trials, k, $< s_k^i >$ for T_s. The decision result, s, is a binary variable, i.e. $s = 1$, if T_s is waived; $s = 0$, if T_s is chosen. To perform the fit, we used the firing activity within a 100 ms time window before T_s onset in FD task (**Figure 4.6A**) for R_A and R_B:

$$cc^i = 1 - P_{sure}(|R_A^i - R_B^i|) = b_1 + \frac{a}{1 + \exp(k|R_A^i - R_B^i| - b_0)} \tag{4.7}$$

Using all trials in FD task, we obtained $b_0 = 2.22$ Hz; $b_1 = 1.01$; $a = -1.01$; $k = 0.089$ ($R^2 = 0.98$, **Figure 4.6A**, red dash line). Importantly, a real result here is to quantify confidence as a function of the neural activity. Confidence estimation is thus applicable to all trials, even without sure target presentation. We then used these

estimated parameters to calculate cc^i for each sample trial in both FD and RT tasks, where R_A and R_B are the average firing rates within a 100 ms time window before T_s onset in FD task (**Figures 4.6, 4.7A, and 4.8C**) and those before one of the bumps reaching a decision threshold in RT task (**Figure 4.9**).

Chapter 5

Conclusions and future directions

A recurring theme throughout this dissertation is the emphasis on model-based statistical analysis of neural recording data. This yields a deeper understanding of the emerging role of computational neuroscience in interpreting and designing population recording experiments (Brown et al., 2004; Stevenson and Kording, 2011; Cunningham and Yu, 2014; Freeman et al., 2014; Freeman, 2015; Harris et al., 2016; Aljadeff et al., 2016; Ji et al., 2016). This dissertation reveals two basic challenges in current studies of simultaneously recorded neural dynamics and discusses solutions based on data collected from a decision-making task (where the neural activity exhibited complicated dynamics). The first challenge deals with a potential problem with modern data collection. Laboratories often examine different properties of neural dynamics in the same local neural circuit using different recording technologies. How does one summarize and compare neural dynamics from these similar yet distinguish-

able datasets? The other problem initiates a general discussion of the necessity of simultaneous recording in neural dynamics research. This could be answered without doubt, but lacks quantitative measurements.

Advances in recording technologies enable us to determine the neural dynamics at different spatiotemporal resolutions. Nevertheless, each technology has advantages and disadvantages, as well as innate limitations. This is the impetus for the tendency of laboratories to examine the same neural circuit by combining recording technologies, for instance electrophysiology and calcium imaging. Despite the fact that such integration of methodologies can significantly quicken the pace of scientific discovery in neuroscience, it still remains a crucial challenge for us to determine how to merge these data together to tell a consistent story about neural circuits. Therefore, an emerging field in computational neuroscience deals with establishing and confirming the phenomenological or neurophysiological models that connect the neural dynamics in different recordings. In **Chapter 2**, we performed a pioneering study in a challenging situation (where the dynamics of the neural circuit were rich and variable across neurons) highlighting this issue, and found that a phenomenological computational model can link spike events in electrophysiology to fluorescence changes in calcium imaging in an informative way. We analyzed electrophysiology and calcium imaging measured in matched neuronal populations from the same decision-making task. We directly compared the results of standard measurements of selectivity and population dynamics. We detected quantitative and qualitative discrepancies at both the level

of single cells and neural populations. Additionally, we discovered that these discrepancies can be explained by the spike-to-calcium model. Therefore, we feel that these biases could be better addressed in future studies of techniques to reverse engineer electrophysiology from imaging, using statistical or other approaches.

Calcium imaging has specific advantages, such as genetic targeting and chronic stability in recordings, and has gradually replaced electrophysiology in certain fields of neuroscience. More advances can be expected in the near future that will improve the temporal resolution and signal-to-noise ratios of calcium imaging recordings, and more studies in neuroscience will be done using calcium imaging. Calcium imaging, however, is inherently an indirect and non-linear reporter of neural activity, and one has to acknowledge that fundamental discrepancies remain within the neural dynamics inferred from calcium imaging. Financially and technically, it is difficult for single laboratories to have the power and resources to implement electrophysiology and calcium imaging at the same time. Computational neuroscientists should play a role in guiding the research across laboratories and helping laboratories using imaging to "translate" their results to electrophysiology.

One future direction of our work is to develop a statistical approach to reverse engineer electrophysiology from imaging and more importantly, to determine the confidence interval of a reverse engineered approach. The reverse engineering approach of electrophysiology from imaging (calcium-to-spike) has been developed for nearly a decade (Yaksi and Friedrich, 2006; Greenberg et al., 2008; Sasaki et al., 2008;

Vogelstein et al., 2009; Vogelstein et al., 2010; Grewe et al., 2010; Oñativia et al., 2013; Park et al., 2013; Pnevmatikakis et al., 2014a; Pnevmatikakis et al., 2014b; Pnevmatikakis et al., 2016; Yang et al., 2016; Ganmor et al., 2016; Theis et al., 2016). It is not until recently that the research has begun to determine the performance of calcium-to-spike models systematically (Theis et al., 2016). The performance in such research was based either on artificial data or on the simultaneous electrophysiology and imaging of the cells at low firing rates. However, the neural dynamics examined in these studies are only a subset of those found in neural data. For instance, neurons at low firing rates usually behave sparsely and have relatively simple dynamics, which can be modeled as a Dirac delta function of spike times. In this case, the number of action potentials detected by complicated calcium-to-spike models improves only a little compared to that using a direct deconvolution of the calcium signal (or even a direct thresholding algorithm). The performance is thus only a weak function of the complexity of the indicator's dynamics or even that of neural dynamics. Our work in this dissertation creates a new path for discussion, where neural dynamics are complicated by their higher spike rates, dynamic response patterns and the variable nature of the innate dynamics of the calcium indicator. In this case, computing true neural dynamics from the data recorded by calcium imaging can be difficult, since the observed neural dynamics could stem from a set of noisy combinations of neural dynamics and innate dynamics of the calcium indicator. Specifically, we propose a few possible directions for future studies of this topic.

Firstly, we propose a new statistical goal of reverse engineering electrophysiology from imaging to find all possible combinations of neural dynamics and the innate dynamics of the calcium indicator, that could result in the observed images. One should not only produce one specific decoding result, but instead provide the confidence intervals for each possible decoding result. This is expected to be done with novel concepts and developments in statistical models (versus a simple maximum likelihood model). Secondly, the newly developed computational models can only be validated with real neural data in a matched dynamical regime, and there remains the technical difficulty of collecting such a dataset for neurons in the frontal cortex. This is a future direction for both experimental and modeling advances. In our collaboration with the laboratory of Dr. Karel Svoboda, we have planned to examine the electrophysiology and simultaneous imaging of neurons in the anterior lateral motor cortex, where the neurons exhibit complicated and variable dynamics during a decision-making task. The validation will be done in two ways: (1) examining whether the ground truth is in the set of possible solutions and then defining a similarity metric that limits the possible decoded neural dynamics to those with high similarity to the ground truth; (2) determining from computational models whether any experimental manipulations can be done to eliminate solutions far from the ground truth, and validating this prediction using experiments.

Due to the nature of the indirect recording, we would expect high uncertainty in neural data. It is thus worth the effort to advance the statistical models to interpret

these neural data. Overall, as the experimental tools and computational models are developed, computational scientists will interact with the experimentalists and move towards understanding of neural data in an optical-imaging era.

Advances in simultaneous recording will affect two major areas of computational neuroscience. As the number of simultaneously recorded neurons increases, one can directly compare large-scale neural simulations with matched large-scale neural recordings. Secondly, more coarse-grained models of network dynamics and population codes will be able to draw from increasingly complete neural data. Accordingly, tools for statistical inference and data analysis should be provided for linking the neural recording to the computational models. In line with this idea, **Chapter 3** shows an early attempt to uncover and extract the computational principles of decision making from simultaneous neural data. Our analysis of simultaneously recorded neural data adopted the idea of a latent variable model, which can simultaneously perform dimensionality reduction and time series analysis. We identified neural dynamics in the latent space, where each neural mode represents a source of input shared by multiple neurons, which could present a continuous neural signal underlying the internal states in different stages of a decision (such as pre-sample, sample, delay, and response) at the population level. Moreover, we found that neural-mode dynamics in this shared-activity space could predict different aspects of behavioral variability, such as trial type, reaction time, and trial correctness, in single trials, and be robustly maintained for seconds post-decision, thus representing single-trial correlates of these properties.

Following this direction, more statistical tools will be developed in the future to enable our ability to validate the computational principles in neural circuits. Thereafter, one can further propose new computational principles by collecting large-scale neural data and developing new statistical tools for model validation. This process will help us draw a clearer picture of neural computations in the intact brain.

Multi-electrode and optical imaging recordings provide simultaneous monitoring of activity from tens to hundreds of neurons, and thus enable our ability to probe the statistical structures of neural population activity. Our ability to visualize neural dynamics is limited to a handful of dimensions (2 - 4 dimensions). We therefore require state-of-the-art statistical tools to help us explore the informative neural dynamics in high dimensional space. Importantly, we emphasize "informative" as the key in search of dimensionality reductions. Unfortunately, there is no simple formula to determine that. Nevertheless, one can choose from practical exemplary analyses that will best uncover certain dynamics of neural data as we examine and compare more models using real data. Notably, the majority of latent variable models, until now, served to analyze electrophysiological data. We thus propose a future direction of model based analysis for large-scale neural data recorded under different methods, each of which will have a suitable and individualized analysis.

Although much work still remains to achieve a complete understanding of the neural dynamics in large-scale neural data, this dissertation provides a stepping stone towards this goal. This research was conducted at an opportune time in the advanc-

ing field of neuroscience, as there is currently much excitement in building statistical models for understanding large-scale neural data and large-scale computational models for extracting the principles of these neural dynamics. This movement is fueled by enthusiastic collaborations between neurophysiologists and theoretical scientists, combining big biological data at various levels with rapidly advancing statistical and computational techniques to answer difficult questions. This dissertation is one fruit of such productive collaborative efforts and we anticipate much more to come.

Appendix A

General Methods in population analysis in neural data

Multi-electrode and optical imaging recordings provide simultaneous monitoring of activity from tens to hundreds of neurons, and thus enable our ability to probe the statistical structure of neural population activity. To exploit this, we need state-of-the-art statistical methods.

Dimensionality reduction (particularly by linear methods) is a modern way to handle the visualization, interpretation and analyses of high dimensional data. Different linear dimensionality reductions aim to capture data perspectives of interest, such as covariance, dynamical structure, margins between data classes, and so on. This chapter will illustrate the details of two modern statistical models heavily used in **Chapter 2** and **3**. One serves the purpose of identifying the coding direction of trial types;

this is a sparse version of linear discriminant analysis (Cunningham and Yu, 2014; Guo et al., 2007). The other infers a linear dynamical system (Ghahramani and Hinton, 1996a) description of the neural network from the population activity of simultaneously recorded neurons, with a variation where the interaction among neurons could be time-dependent (Petreska et al., 2011).

A.1 Sparse linear discriminant analysis

A class of problems in statistics is based on labelled subgroups. Linear discriminant analysis (LDA) is often used as an analysis in such a classification problem (Fisher, 1936; Rao, 1948; Fukunaga, 1990; Krzanowski, 2000; Bishop, 2006; Seber, 2009). The purpose of LDA is to project the data in such a way that separation between subgroups is maximized.

In system neuroscience, LDA is usually applied to infer population neural code for binary choices, or those with a few options (Briggman et al., 2005; Averbeck et al., 2006; Durstewitz et al., 2010; Cunningham and Yu, 2014; Kiani et al., 2015; Li et al., 2016) (e.g. four behavioral epochs in **Figures 2.7G-I**).

Let us take two neuronal activities in the delayed discrimination task (**Figures 2.1A**) for example. In **Figure A1.1A**, two neurons in the task exhibit complex dynamics of selectivity, i.e. one neuron is a monophasic- (Cell #1) and the other is a multiphasic-selective (Cell #2) neuron. One can determine the decision bound

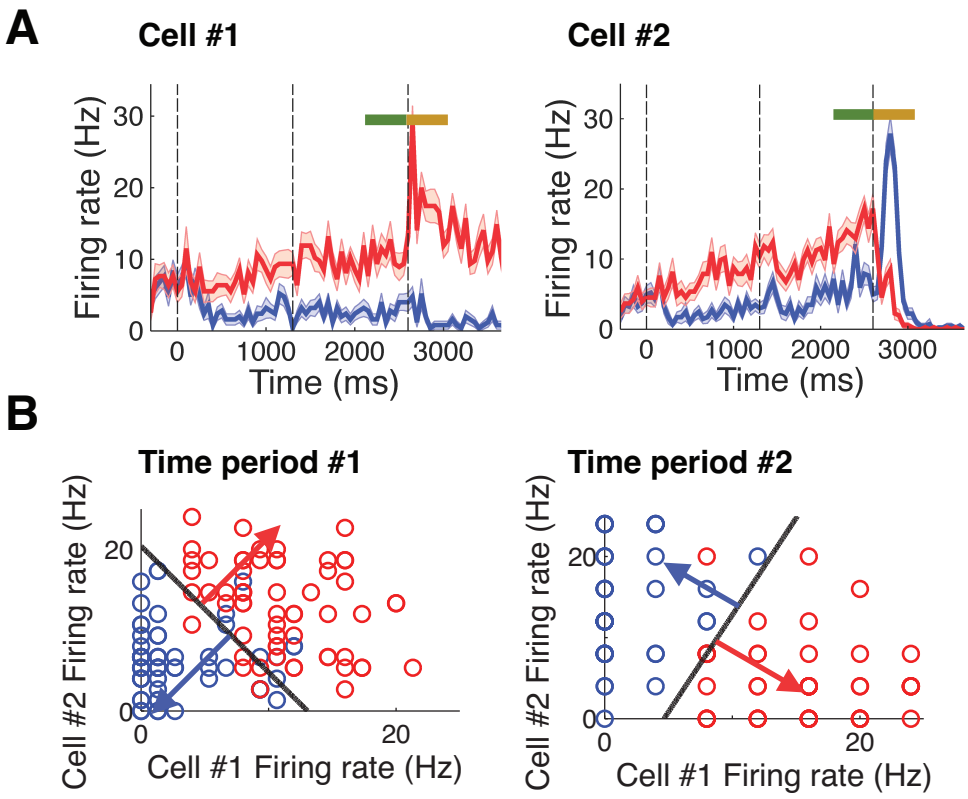


Figure A1.1: Schematic description of linear discriminant analysis for neural code of trial type.
(**A**) Spike activities of two exemplary neurons in the delayed discrimination task (description of the task shown in **Figure 2.1**; blue, trial type A; red, trial type B; thick line, mean activity across trials; shaded area, s.e.m. of the mean activity across trials). The activities were aligned to onset time of the sample period. Cell #1, monophasic selective neuron, left panel; Cell #2, multiphasic selective neuron (which has a switch of selectivity at response), right panel. The green bar marks Time Period #1 (late delay period, 500 ms time window); the yellow bar marks Time Period #2 (early response period, 500 ms time window). (**B**) Decision boundary between trial types A and B in decision space spanned by activities of Cells #1 and #2 at two time periods (Time Period #1, late delay, left panel; Time Period #2, early response, right panel). The decision bounds between activities of two trial types are indicated by black lines (blue, trial type A; red, trial type B); the arrow indicates the optimal LDA decoder for each trial type.

that separates trial type A and B from a combination of activities of Cell #1 and #2 (a so-called population activity vector) in a joint activity space spanned by both. Particularly, the LDA decoder finds the optimal linear decision bound among choices. If a choice is determined by a p-dimensional population vector (p is the number of neurons used in LDA), the linear decision bound is a (p-1)-dimensional hyperplane.

For example, in the late delay period (marked as the green bar; **Figure A1.1B**, left), both neurons show low-firing activities for trial type A (in blue), and the decision bound for trial type A is therefore defined as both activities of Cell #1 and #2 being low. The population activity (combinations of the activities of Cell #1 and #2) on the left of the joint activity space encodes trial type A, while that on the right encodes trial type B. Similar analysis applies to the population activity in the early response period of task, except that Cell #2 exhibits a switch of selectivity, which changes the decision bound from the late delay to the early response period of task (**Figure A1.1B**, right).

In the following two sections, we will discuss the basic mathematical model of LDA (**Section A.1.1**) and its robust implementation through regularization that can handle the inversion of the covariance matrix even when it is ill-conditioned (**Section A.1.2**).

A.1.1 Description of linear discriminant analysis

Let us assume that there are G different labels encoded by a population of neurons ($\mathbf{r} \in \mathbf{R}^p$, p is the number of the neurons). Each label represents a choice, which could be a combination of several behavioral parameters. For example, a label could be a correct type-A trial, which is a product of reward (correct or error) and stimulus conditions (A or B).

Moreover, we assume that the subset of the population activity in each label, $g \in \{1, \cdots, G\}$ has a multivariate normal distribution with a common covariance matrix $\mathbf{\Sigma}$ of dimension $p \times p$ and mean vectors μ_g.

Suppose we now have a random sample of $n = \sum_{g=1}^{G} n_g$ observations from these labels with their true group labels being unknown, where n_g stands for the number of observations in Label $\#g$. Our question is then how to correctly identify the label, to which an observation belongs. Explicitly, we defined $\mathbf{r}_{g,i}$ as the ith observation in Label Group $\#g$ and

$$\mathbf{r}_{g,i} \sim \mathrm{N}(\mu_g, \mathbf{\Sigma}),$$

where $i \in 1, \cdots, n_g$.

We now use LDA to classify observation $\mathbf{r}_{g,i}$ to a label $\tilde{g}$, which minimizes its neuronal projection on to the mean of label $\tilde{g}$, $(\mathbf{r}_{g,i} - \mu_{\tilde{g}})^T \mathbf{\Sigma}^{-1} (\mathbf{r}_{g,i} - \mu_{\tilde{g}})$

$$\tilde{g} = \arg\min_{g'} (\mathbf{r}_{g,i} - \mu_{g'})^T \mathbf{\Sigma}^{-1} (\mathbf{r}_{g,i} - \mu_{g'}).$$

Alternatively, this is equivalent to finding the label that maximizes the likelihood of the observation. More often, one would have some prior knowledge as to the proportion of each label (e.g. discrete uniform distribution is usually applied). For example, let π_g be the proportion of Label $\#g$ such that $\sum_{g=1}^{G} \pi_g = 1$. Then, instead of maximizing the likelihood, we maximize the posterior probability, $P(g|\mathbf{r}_{g,i}) \propto P(\mathbf{r}_{g,i}|g)P(g)$, the observation belongs to a particular label, i.e.

$$\tilde{g} = \arg\max_{g'} \pi_{g'} \exp[-\frac{1}{2}(\mathbf{r}_{g,i} - \mu_{g'})^T \mathbf{\Sigma}^{-1}(\mathbf{r}_{g,i} - \mu_{g'})].$$

The linearity of LDA method comes from the assumption of common covariance matrix $\mathbf{\Sigma}$, which simplifies the above criterion as

$$\tilde{g} = \arg\min_{g'} \left\{ \mathbf{r}_{g,i}^T \mathbf{\Sigma}^{-1} \mu_{g'} - \frac{1}{2} \mu_{g'}^T \mathbf{\Sigma}^{-1} \mu_{g'} + \log \pi_{g'} \right\},$$

which is the so-called discriminant function of LDA.

In reality, experimenters would use LDA as a type of supervised learning, where both the centroid of Label $\#g$, μ_g, and, the common covariance matrix, $\mathbf{\Sigma}$, are unknown but estimated from the samples in a training set,

$$\hat{\mu}_g = < \mathbf{r}_{g,i} >_i = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{r}_{g,i},$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}(\mathbf{R} - < \mathbf{R} >)(\mathbf{R} - < \mathbf{R} >)^T,$$

where $\mathbf{R}$ is a $p \times n$ matrix with each column corresponding to an observation in population activity $\mathbf{r}$, and $< \mathbf{R} >$ is a matrix of the same dimensions with each column corresponding to the sample mean vector of the label that the column belongs to.

Notably, when the assumption of the common covariance matrix is not satisfied, one could use an individual covariance matrix for each group and this leads to the so-called quadratic discriminant analysis where the discriminating boundaries are quadratic curves(Krzanowski, 2000; Seber, 2009).

A.1.2 Sparse linear discriminant analysis

Neurons in the frontal cortex usually fire at a wide range of rates. Taking anterior lateral motor cortex (ALM) neurons for example **Figure 2.S2D**), we find that their firing rates range from zero to > 50 Hz. Poisson distributions, $Poiss(r)$, are considered to be good estimates of neuronal activities (Dean, 1981; Tolhurst et al., 1983; Bradley et al., 1987; Scobey and Gabor, 1989; Vogels et al., 1989; Snowden et al., 1992; Britten et al., 1993; Softky and Koch, 1993; Geisler and Albrecht, 1997; Shadlen and Newsome, 1998; Churchland et al., 2011) and its variability across trials (where r is the firing rate of the neuron), and can be approximated to a Gaussian distribution, $N(r, r)$ for neurons that spike at high rates, but not those that spike only at a few Hz. Unfortunately, the LDA decoder and its performance are bounded by the limitation of the assumption of Gaussianity of covariance among the neuronal spike rates. One

should keep in mind the need to exclude the use of neurons that fire at low rates.

Moreover, it is often difficult to estimate the optimal LDA decoder in high-dimensional neural space (more than 1500 units in our study), despite the simplicity of its mathematical form, for two major reasons. First, there could be a null space spanned by a considerable fraction of neurons (Druckmann and Chklovskii, 2010; Druckmann and Chklovskii, 2012; Kaufman et al., 2014). This leads to a degeneracy or singularity of covariance matrix, which then cannot be inverted. This problem is more likely to occur if low firing rate neurons are included in the analysis. In this case, although we may use a generalized version of matrix inversion, the estimation will be unstable. Secondly, the need for matrix operations with high-dimensionality hinders the applicability of LDA.

Therefore, we want to search for a version of LDA that can (1) limit the use of low firing rate units (those that are in fact statistically non-Gaussian), and (2) correct even in the case of a singular or ill-conditioned covariance matrix, and thus capable of computing a robust optimal decoder. We adopted a sparse LDA algorithm to achieve this (Guo et al., 2007).

To resolve the singularity problem, instead of using $\hat{\mathbf{\Sigma}}$, we use

$$\bar{\mathbf{\Sigma}} = \alpha\hat{\mathbf{\Sigma}} + (1-\alpha)\mathbf{I}_p,$$

where regularization parameter $0 \leq \alpha \leq 1$ (Hoerl and Kennard, 1970). One could

further shrink centroids using a l_1 regularization (Friedman et al., 2001; Tibshirani et al., 2002), where the shrunken centroids are computed using

$$\hat{\mu}'_g = \text{sign}(\hat{\mu}_g)(|\hat{\mu}_g| - \Delta)_+,$$

where Δ is the l_1 regularizer and two regularization parameters (α, Δ) could be decided using cross-validation.

A.2 Time-varying linear dynamical system analysis

Modern statistical analyses provide unprecedented insights into the structure of neural population activity in a high dimensional space. One approach uses dimensionality reduction methods (Brenner et al., 2000), such as principal component analysis (Jolliffe, 2014) and factor analysis. Furthermore, for the simultaneous recordings of neural data, one would also apply or combine the temporal analysis to identify the simultaneous population activity and link it to external stimuli and observed behavior.

Two possible classes of models, a class of supervised learning models (namely, discriminative models) and a class of unsupervised learning models (namely, generative models), can be used to exploit the spatiotemporal dynamics of population activity

that have become available through multi-neuron recording methods.

A generalized linear model (GLM) follows supervised learning. In GLM, the model applies external stimuli and spiking history as covariates driving the spiking of the neural population (Paninski, 2004; Truccolo et al., 2005; Pillow et al., 2008; Vidne et al., 2012). The interdependence of different neurons is modeled by terms that link the instantaneous firing rate of each neuron to the recent spiking history of the population. The parameters of the GLM can be learned efficiently by convex optimization (Chornoboy et al., 1988; McCullagh and Nelder, 1989; Pillow et al., 2008; Vidne et al., 2012). Such models have been successful in a range of studies of population recordings (Pillow et al., 2008; Vidne et al., 2012).

The latent-variable model (Everitt, 1984; Bishop, 1998; Knott and Bartholomew, 1999) is an alternative approach, resembling unsupervised learning. These latent-variable based models, e.g. Gaussian Process Factor Analysis (GPFA) (Yu et al., 2009) and the other state-space models (Roweis and Ghahramani, 1999; Smith and Brown, 2003; Lawhern et al., 2010; Macke et al., 2011; Petreska et al., 2011; Pfau et al., 2013; Buesing et al., 2014), are an extension of factor analysis, where models share variability (off-diagonal elements in the correlation matrix across neurons). In turn, such shared variability is considered to be driven by sources of common inputs (Kulkarni and Paninski, 2007; Paninski et al., 2010; Vidne et al., 2012). These analyses have been used to extract low-dimensional hidden structure that captures the variability of the recorded data, both in time and across the population of neurons.

Furthermore, the extracted low-dimensional hidden structures can be used to visualize population activity, and be linked to the observed behaviors (Afshar et al., 2011; Churchland et al., 2012; Gilja et al., 2012; Kao et al., 2015; Kaufman et al., 2014; Kaufman et al., 2015; Mante et al., 2013; Sussillo et al., 2015).

A.2.1 Description of linear dynamical system analysis

Linear dynamical systems (LDS) models are a modern technique (Shumway and Stoffer, 1982; Ghahramani and Hinton, 1996a), within the class of GPFA (Yu et al., 2009). They implement a continuous form of the Hidden Markov Model (Rabiner and Juang, 1986; Ghahramani and Hinton, 1996b), and are able to perform simultaneous analysis of temporal dynamics and dimensionality reduction (Cunningham and Yu, 2014). They were first proposed and used in the field of physics and engineering (Zhou et al., 1996; Ljung, 1998; Verhaegen and Verdult, 2007). Recently, neuroscientists also started applying these models in analyses of simultaneous recordings of population activity and used it as an engineering tool for translational applications of brain-computer interfaces (Gilja et al., 2012; Kao et al., 2015).

LDS comes with simple mathematical equations. Imagine that we have n neurons $\mathbf{r} \in \mathbf{R}^n$ being simultaneously recorded during a behavioral task (**Figure A1.2A**). The dynamics of neurons exhibit a strong correlation in time; we could therefore perform a

dimensionality reduction to summarize dynamics of a subset of strongly coordinated neurons to latent modes $\mathbf{x} \in \mathbf{R}^m$, where $m < n$ for the purpose of dimensionality reduction.

$$\mathbf{r}_t^k = \mathbf{C}\mathbf{x}_t^k + \mathbf{r}_0 + \mathbf{v}_t^k, \tag{A.1}$$

where superscript k stands for the trial index; $k \in \{1, \cdots, K\}$ and K is the total number of collected trials; subscript t stands for the time index in a trial; $t \in \{1, \cdots, T\}$ and T is the total number of discrete time bins in a recording (the recording length is thus equal to $\Delta_t \cdot T$). $\mathbf{r}_0 =< \mathbf{r}_t^k >_{k,t}$ is the mean activity of the neuron across time and trials; $\mathbf{C}$ is the projection matrix from the latent modes to observed neuronal activities; $\mathbf{v}_t^k$ is an independent noise model for each neuron that parameterizes the distribution of $\mathbf{r}_t^k$ based on the mode and some hyper-parameters. Although firing rates are typically modeled as as a Poisson process $\mathbf{r}_t^k \sim Poiss(\mathbf{C}\mathbf{x}_t^k + \mathbf{r}_0)$, $\mathbf{v}_t^k$ can be approximated as a Gaussian model empirically, $\mathbf{v}_t^k \sim N(0, \mathbf{\Sigma}_v)$ (where $\mathbf{\Sigma}_v$ is a diagonal matrix $\{\sigma_{v_1}^2, \cdots, \sigma_{v_n}^2\}$), at least when the neurons fire at a considerable rate ($> 3Hz$ based on our analyses).

The dynamics of the neurons are then modeled explicitly as those of the latent modes.

$$\mathbf{x}_t^k = \mathbf{A}\mathbf{x}_{t-1}^k + \mathbf{w}_{t-1}^k, \tag{A.2}$$

which is a linear dynamics system or first-order autoregressive process. $\mathbf{A}$ represents the interactions of the latent modes (the off-diagonal elements) and decays of dynam-

ics of each latent mode (the diagonal elements). $\mathbf{w}_t^k$ follows a Gaussian distribution that $\mathbf{w}_t^k \sim N(0, \mathbf{Q})$. It presents a source of variability from time to time and from trial to trial, which is not necessarily independent in time. One could consider the post-hoc estimation of the time series of $\mathbf{w}_t^k$ to be noisy inputs onto the latent modes at Trial #k.

Given that the time index of a latent model runs from zero to T, in order to perform the parameter estimation of LDS, we also need to specify the model of the initial state $\mathbf{x}_1$. Here we take a simple model of the initial state as $\mathbf{x}_1 \sim N(\pi_\mathbf{0}, \mathbf{Q_0})$. $\mathbf{Q_0}$ represents a pre-task state of the system, and could differ from the $\mathbf{Q}$ in the task state (Churchland et al., 2010b). Now we can explicitly write the joint distribution among the neural activity and the latent modes (a discrete operation of the linear system using Hidden Markov Model) as follows:

$$P(\mathbf{x}_{1\cdots T}, \mathbf{r}_{1\cdots T}) = P(\mathbf{x}_1) \prod_{t=2}^{T} P(\mathbf{x}_t|\mathbf{x}_{t-1}) \prod_{t=1}^{T} P(\mathbf{r}_t|\mathbf{x}_t).$$

Our goal is to estimate the parameter set Θ from K time series of observations $\{\mathbf{r}_t\}_k$, where

$$\Theta = \{\mathbf{A}, \mathbf{Q}, \pi_\mathbf{0}, \mathbf{Q_0}, \mathbf{C}, \mathbf{r}_0, \mathbf{R}\}, \tag{A.3}$$

and the joint log probability (the cost function) is then a sum of quadratic terms

$$
\begin{aligned}
L(\Theta; \mathbf{x}^k_{1\cdots T}, \mathbf{r}^k_{1\cdots T}) &= \log P(\mathbf{x}^k_{1\cdots T}, \mathbf{r}^k_{1\cdots T}; \Theta^c) \\
&= \log P(\mathbf{x}^k_1; \Theta^c) \prod_{t=2}^{T} P(\mathbf{x}^k_t | \mathbf{x}^k_{t-1}; \Theta^c) \prod_{t=1}^{T} P(\mathbf{r}^k_t | \mathbf{x}^k_t; \Theta^c) \\
&= -\frac{1}{2} \Bigg\{ T(m+n)\log(2\pi) + \log|\mathbf{Q_0}| + (T-1)\log|\mathbf{Q}| \\
&\quad + T\log|\mathbf{R}| + (\mathbf{x}^k_1 - \pi_\mathbf{0})' {\mathbf{Q_0}}^{-1} (\mathbf{x}^k_1 - \pi_\mathbf{0}) \\
&\quad + \sum_{t=2}^{T} (\mathbf{x}^k_t - \mathbf{A}\mathbf{x}^k_{t-1})' \mathbf{Q}^{-1} (\mathbf{x}^k_t - \mathbf{A}\mathbf{x}^k_{t-1}) \\
&\quad + \sum_{t=1}^{T} (\mathbf{r}^k_t - \mathbf{C}\mathbf{x}^k_t - \mathbf{r}_0)' \mathbf{R}^{-1} (\mathbf{r}^k_t - \mathbf{C}\mathbf{x}^k_t - \mathbf{r}_0) \Bigg\}.
\end{aligned}
$$

The parameter set Θ can be estimated using the Expectation-Maximization Algorithm (**EM**).

One thus needs to compute expectation of $\mathbf{x}^k_t$, $E[\mathbf{x}^k_t | \mathbf{y}^k_{1\cdots T}, \Theta^c]$, its variance, $Var[\mathbf{x}^k_t | \mathbf{y}^k_{1\cdots T}, \Theta^c]$ and its covariance in a one-time step $Cov[\mathbf{x}^k_t, \mathbf{x}^k_{t-1} | \mathbf{y}^k_{1\cdots T}, \Theta^c]$ and these estimations (**E-Step**) can be done using the Kalman forward-backward algorithm. See algorithm block **Algorithm 1** for details.

In **M-step**, one needs to find the parameter to maximizes the quantity:

$$\Theta^* = \arg\max_{\Theta} Q(\Theta | \Theta^c)$$

The pre-processing of the data and some additional assumptions about the pa-

Algorithm 1 Kalman forward-backward algorithm:
$\mathbf{x}_t(k) = E[\mathbf{x}_t^k | \mathbf{y}_{1\cdots T}^k, \Theta^c]$;
$\mathbf{Q}_t(k) = Var[\mathbf{x}_t^k | \mathbf{y}_{1\cdots T}^k, \Theta^c]$;
$\mathbf{Q}_{t,t-1}(k) = Cov[\mathbf{x}_t^k, \mathbf{x}_{t-1}^k | \mathbf{y}_{1\cdots T}^k, \Theta^c]$

for all $k \in \{1 \cdots K\}$ **do**

// **Forward Step**
$\mathbf{x}_1^0(k) \leftarrow \pi_{\mathbf{0}}$, $\mathbf{Q}_1^0(k) \leftarrow \mathbf{Q_0}$
for $t = 1$ to T **do**
$\mathbf{K}_t \leftarrow \mathbf{Q}_t^{t-1}(k)\mathbf{C}'(\mathbf{R} + \mathbf{C}\mathbf{Q}_t^{t-1}(k)\mathbf{C}')^{-1}$
$\mathbf{Q}_t^t(k) \leftarrow \mathbf{Q}_t^{t-1}(k) - \mathbf{K}_t\mathbf{C}\mathbf{Q}_t^{t-1}(k)$
$\mathbf{x}_t^t(k) \leftarrow \mathbf{x}_t^{t-1}(k) + \mathbf{K}_t(\mathbf{y}_t(k) - \mathbf{C}\mathbf{x}_t^{t-1}(k) - \mathbf{r}_0)$
if $t < T$ **then**
$\mathbf{Q}_{t+1}^t(k) \leftarrow \mathbf{A}\mathbf{Q}_t^t(k)\mathbf{A}' + \mathbf{Q}$
$\mathbf{x}_{t+1}^t(k) \leftarrow \mathbf{A}\mathbf{x}_t^t(k) + \mathbf{b_t}$
end if
end for{Forward step}

// **Backward Step**
$\mathbf{x}_T(k) \leftarrow \mathbf{x}_T^T(k)$, $\mathbf{Q}_T(k) \leftarrow \mathbf{Q}_T^T(k)$
$\mathbf{Q}_{T,T-1}(k) \leftarrow (\mathbf{I} - \mathbf{K}_T\mathbf{C})\mathbf{A}\mathbf{Q}_{T-1}^{T-1}(k)$
for $t = T-1$ to 1 step -1 **do**
$\mathbf{J}_t \leftarrow \mathbf{Q}_t^t(k)\mathbf{A}'(\mathbf{Q}_{t+1}^t(k))^{-1}$
$\mathbf{Q}_t(k) \leftarrow \mathbf{Q}_t^t(k) + \mathbf{J}_t(\mathbf{Q}_{t+1}(k) - \mathbf{Q}_{t+1}^t(k))\mathbf{J}_t'$
$\mathbf{x}_t(k) \leftarrow \mathbf{x}_t^t(k) + \mathbf{J}_t(\mathbf{x}_{t+1}(k) - \mathbf{A}\mathbf{x}_t^t(k))$
if $t > 1$ **then**
$\mathbf{Q}_{t,t-1}(k) \leftarrow \mathbf{Q}_t^t(k)\mathbf{J}_{t-1}' + \mathbf{J}_t(\mathbf{Q}_{t+1,t}(k) - A\mathbf{Q}_t^t(k))\mathbf{J}_{t-1}'$
end if
end for{Backward step}

return $\mathbf{x}_t(k), \mathbf{Q}_t(k), \mathbf{Q}_{t,t-1}(k)$

end for{K trials}

rameters are helpful to simplify the parameter estimation. Specifically, we removed the mean activity of neural observations,

$$\mathbf{r}_0 \stackrel{\text{def}}{=} < \mathbf{r}_t^k >_{t,k},$$

where $\frac{1}{TK}\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbf{y}_t^k = \mathbf{0}$. And we further assume that $\frac{1}{TK}\sum_{k=1}^{K}\sum_{t=1}^{T}E[\mathbf{x}_t^k] = \mathbf{0}$. In this case,

$$\begin{aligned}
\pi_{\mathbf{0}}{}^* &= \frac{1}{K}\sum_{k=1}^{K}\left(E[\mathbf{x}_1^k]\right) \\
\mathbf{Q_0}{}^* &= \frac{1}{K}\sum_{k=1}^{K}\left((E[\mathbf{x}_1^k] - \pi_{\mathbf{0}}{}^{new})(E[\mathbf{x}_1^k] - \pi_{\mathbf{0}}{}^{new})' + Var[\mathbf{x}_1^k]\right) \\
\mathbf{C}^* &= \left(\sum_{k=1}^{K}\sum_{t=1}^{T}\mathbf{y}_t^k E[\mathbf{x}_t^k]'\right)\left(\sum_{k=1}^{K}\sum_{t=1}^{T}(E[\mathbf{x}_t^k]E[\mathbf{x}_t^k]' + Var[\mathbf{x}_t^k])\right)^{-1} \\
\mathbf{R}^* &= \frac{1}{TK}\sum_{k=1}^{K}\sum_{t=1}^{T}\left((\mathbf{y}_t^k - \mathbf{C}^* E[\mathbf{x}_t^k])\mathbf{y}_t^{k\prime}\right) \\
\mathbf{A}^* &= \left(\sum_{k=1}^{K}\sum_{t=2}^{T}(Cov[\mathbf{x}_t^k, \mathbf{x}_{t-1}^k] + E[\mathbf{x}_t^k]E[\mathbf{x}_{t-1}^k]')\right) \\
&\quad \times\left(\sum_{k=1}^{K}\sum_{t=2}^{T}(Var[\mathbf{x}_{t-1}^k] + E[\mathbf{x}_{t-1}^k]E[\mathbf{x}_{t-1}^k]')\right)^{-1} \\
\mathbf{Q}^* &= \frac{1}{K(T-1)}\sum_{k=1}^{K}\sum_{t=2}^{T}\Bigg(E[\mathbf{x}_t^k]E[\mathbf{x}_t^k]' + Var[\mathbf{x}_t^k] \\
&\quad -\mathbf{A}^*(E[\mathbf{x}_{t-1}^k]E[\mathbf{x}_t^k] + Cov[\mathbf{x}_{t-1}^k, \mathbf{x}_t^k])\Bigg).
\end{aligned}$$

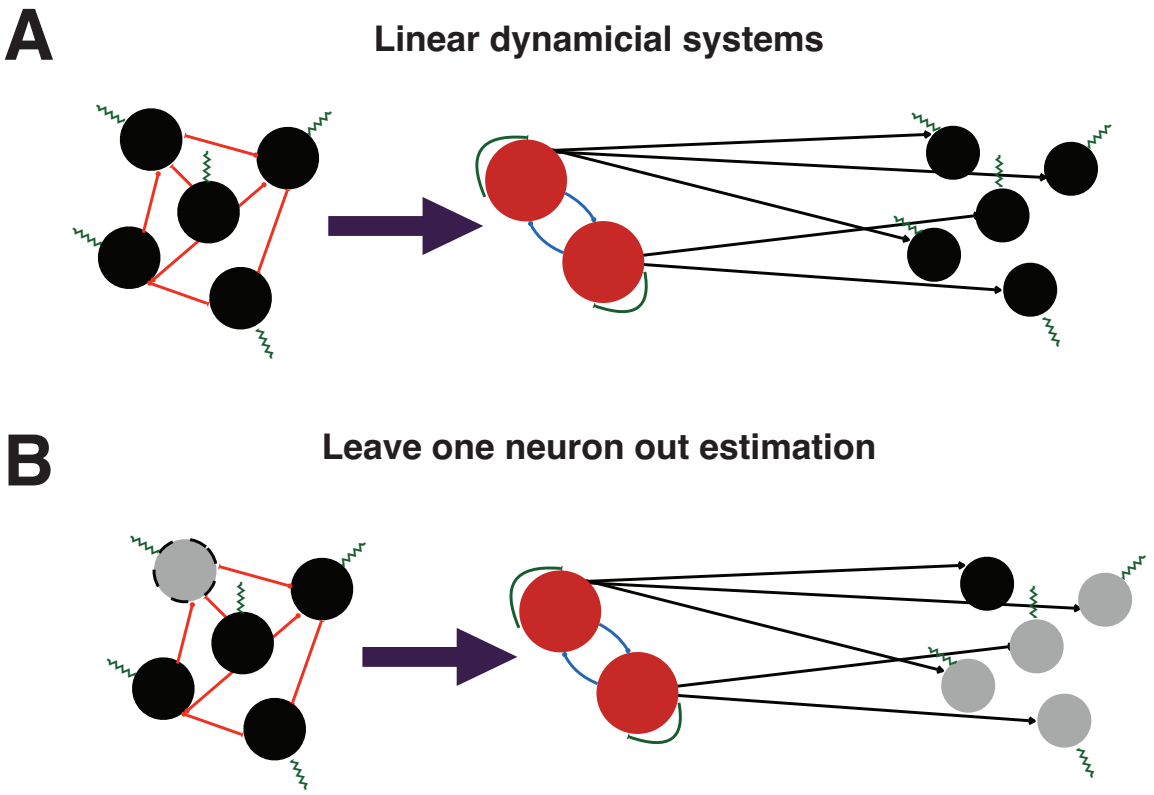


Figure A1.2: Schematic description of linear dynamical system analysis and leave-one-neuron-out estimation.
(**A**) Schematic description of linear dynamical system analysis. Dynamics of highly correlated neurons can be represented as two-layer network, i.e. a latent network (red) usually in a low-dimensional space (spanned by neural modes) and a projection network, which is compared directly to the original neural space. The latent network was assumed to model the shared input across the neurons, which drive the dynamics of the correlations. Each mode in the latent dynamics could be a source of the shared input and thus drive the common dynamics of a subpopulation of chorusing neurons. In the projection space (i.e. original neural space), each neuron could receive shared input from several modes and also a source of independent input (green zigzag lines), which is shared across neurons. (**B**) Schematic description of leave-one-neuron-out estimation. The dynamics of a neuron (grey circle with black dash line) that shared a common source of input from modes can be fitted from the dynamics of other neurons (black circles) whose activities have strong correlations with it.

A.2.2 Leave one neuron out estimation

Leave-one-neuron-out (LONO) method is a modern way of determining the goodness-of-fit for the latent-space models as applied to the neuronal data (Yu et al., 2009). It is a general method adopted from cross-validation and a time-efficient version of leave-p-neurons-out ($p \in \{1, 2, \cdots, \frac{n}{2}\}$, where $p = 1$ and n is the number of neurons in a model) with a limited computation time (in general, the leave-p-neuron-out requires a combinatorial computation time $C\binom{n}{p}$ to go through all possible leave-out scenarios) (Fukunaga, 1990; Coyer, 2014; Bishop, 2006).

Active neurons in the cortex usually exhibit a strong correlation across units and time, coordinating with a specific behavior (Gilja et al., 2012; Sussillo et al., 2015; Mante et al., 2013; Kaufman et al., 2014; Ames et al., 2014; Churchland et al., 2012; Kao et al., 2015; Afshar et al., 2011; Kaufman et al., 2015; Ahrens et al., 2012; Peron et al., 2015; Vladimirov et al., 2014; Li et al., 2016). Dynamics of these highly correlated neurons can be modeled using LDS (**Figure A1.2A**) (Kao et al., 2015). Specifically, LDS assumes that the correlated dynamics can be represented explicitly in a low-dimensional space, spanned by a few neuronal modes (that are the population activities across neurons). Each neuron mode (red circles, **Figure A1.2A**) presents a source of shared input across neurons in a local circuit during the recordings. Such a shared input could be a representation of some recurrent inputs in the local circuit or a common-source input outside the local circuit, which drives the common dynamics of a subpopulation of chorusing neurons. Modes could have interactions in time, and

their dynamics (green lines, **Figure A1.2A**) and interactions (blue lines, **Figure A1.2A**) could be modeled using a linear dynamical system (see **Equation A.2**). In this way, one could estimate the dynamics of the shared inputs (the neural modes) using LDS following the **EM** algorithm. To compare directly with the observed neural dynamics recorded, one needs to project the dynamics of the modes back to the original neural space (also called the "projection space"). Notably, the dynamics of neurons in the projection space are composed of two classes of input. One is the direct projection input from the shared modes and the other is the independent input (green zigzag lines, **Figure A1.2A**) from some unknown sources, each of which only drives the variability in dynamics of a single neuron.

One can thus estimate the network structure in a subset of simultaneous recordings (a training dataset; K trials in $\{1, \cdots, K\}$) using LDS and then determine the performance of LDS using the remaining recordings (a testing dataset; L trials in $\{K+1, \cdots, K+L\}$). This procedure is called cross-validation in statistical learning for a generative model (like LDS) (Stone, 1977; Shibata, 1989; Watanabe, 2010). Particularly, if one neuronal activity is driven by a specific neural mode in LDS, its dynamics can be estimated using the other neurons driven by the same mode. To perform the leave-one-neuron-out estimation, one should first estimate the dynamics of modes from the $n-1$ population with a known two-layer network structure, Θ

(**Equation A.3**), from the training dataset,

$$E(\mathbf{x}_t^l|\mathbf{r}^l_{-j,(1,\cdots,T)}, \Theta^K_{-j}),$$

where the jth neuron is eliminated from estimation, and Θ^K is the parameter set estimated from the training dataset $\mathbf{r}_t^k$, $k \in \{1, \cdots, K\}$. One should then compute the estimation of the jth neural dynamics through projections $\mathbf{C}_j^K$,

$$\hat{r}^l_{j,t} = \mathbf{C}_j^K E(\mathbf{x}_t^l|\mathbf{r}^l_{-j,(1,\cdots,T)}, \Theta^K_{-j}) + r_{j,0}, \tag{A.4}$$

where $r_{j,0}$ is the mean activity of Neuron #j.

The performance of an LDS model is thus evaluated as the estimation error between the neural dynamics in the testing dataset and its estimation,

$$R^2 = 1 - \frac{\sum_{j=1}^n < r^l_{j,t} - \hat{r}^l_{j,t} >_{l,t}}{\sum_{j=1}^n < r^l_{j,t} - r_{j,0} >_{l,t}}$$

which is a measure of explained variance for LDS fit in neural dynamics (Fukunaga, 1990; Coyer, 2014; Bishop, 2006).

A.2.3 Time-varying latent dynamical system analysis

In the previous section, we described the model of linear dynamical systems and the estimation of its parameter set using EM algorithm. As we mentioned briefly above, the neural variability, e.g. $\mathbf{Q}$, could differ across states of behaviors (Churchland et al., 2010b). For example, $\mathbf{Q}_0$ at Time "One" (initial state) is not necessarily equal to $\mathbf{Q}$ in the remaining trial. Here we try to explore a more general case that the neural variability was modeled explicitly as a function of time, and for the sake of simplicity, we assume that $\mathbf{Q}$ is a function of behavior epochs (e.g. presample, sample, delay and response in **Figure 2.1**). We named this analysis the time-varying latent dynamical system analysis (TLDS).

The description of TLDS is identical to that of LDS (see **Equations A.1** and **A.2**), except that parameters $\mathbf{A}$, $\mathbf{Q}$, $\mathbf{C}$, $\mathbf{R}$ are behavioral state dependent (**Equations 3.1** and **3.2**). For simplicity, we assumed that they exclusively depended on behavioral epochs. The estimation of parameters follows the same **EM** procedure as that in LDS (**Equations 3.3**, **3.4**, **3.5** and **3.6**).

In this model, we explicitly assumed that the variability changes at the moment of behavioral transitions (see **Methods** in **Chapter 3**). This "instantaneous-switch" assumption holds true for the neural dynamics of ALM neurons. For example, the coding direction of trial type shows a transition change at the switch from delay to

response epochs in electrophysiological data (Li et al., 2016). Nevertheless, some subpopulation of neurons could show a delayed transition of neural dynamics compared to the behavioral state (Petreska et al., 2011). Alternatively, one could exploit the Hidden Markov Model to estimate the transitions of neural states from simultaneous neural recordings implicitly (Jones et al., 2007; Cohen and Maunsell, 2010; Petreska et al., 2011; Bollimunta et al., 2012; Afshar et al., 2011; Latimer et al., 2015; Morcos and Harvey, 2016). Recent studies showed that some neural transitions could be correlated with those of behavioral states (Petreska et al., 2011). This demonstrates that some internal state changes in neural activity cannot be observed at the behavioral level.

A.2.4 Performance of time-varying latent dynamical system analysis on neural data

In this section, we will discuss the TLDS model fit for neural data and its validation based on the LONO estimation.

We start our examination of performance using artificial data. For artificial data, we simply follow the generation of neural dynamics using the exact form of **Equations A.1** and **A.2** for a given set of Θ, even though Gaussian variability is usually not observed in neural dynamics. Alternatively, one could choose a more realistic model to replace **Equation A.1** in generation of artificial data. For example, a Poisson-like

process (Dean, 1981; Tolhurst et al., 1983; Bradley et al., 1987; Scobey and Gabor, 1989; Vogels et al., 1989; Snowden et al., 1992; Britten et al., 1993; Softky and Koch, 1993; Geisler and Albrecht, 1997; Shadlen and Newsome, 1998; Churchland et al., 2011) or a Lognormal-like process (Rupasov et al., 2012; Zylberberg et al., 2011; Mizuseki and Buzsáki, 2013; Buzsáki and Mizuseki, 2014; Buzsáki, 2015) would be more realistic for generation of spike counts in a small discrete time bin. In this case, we apply the following Poisson-like model, which was used predominantly in GLM models (Park et al., 2014; Pillow et al., 2008; Vidne et al., 2012) or Poisson LDS models (Macke et al., 2011; Zhao and Park, 2016),

$$\mathbf{r}_t^k = Poiss(\exp(\mathbf{C}^K \mathbf{x}_t^k + \mathbf{r}_0)), \tag{A.5}$$

where the mean firing rate is a nonlinear function of latent neural modes, $\exp(\mathbf{C}^K \mathbf{x}_t^k + \mathbf{r}_0)$.

For the sake of simplicity, we first evaluate the performance of TLDS fit on artificial data using Gaussian variability in the "projection space". The data was generated using **Equations A.1** and **A.2**, where the parameter set of Θ was given. In this case, one could compute directly the noiseless version of $\mathbf{r}_t^k$,

$$\bar{\mathbf{r}}_t^k = \mathbf{C}\mathbf{x}_t^k + \mathbf{r}_0, \tag{A.6}$$

where the independent noise for each single neuron was removed, and then compare it

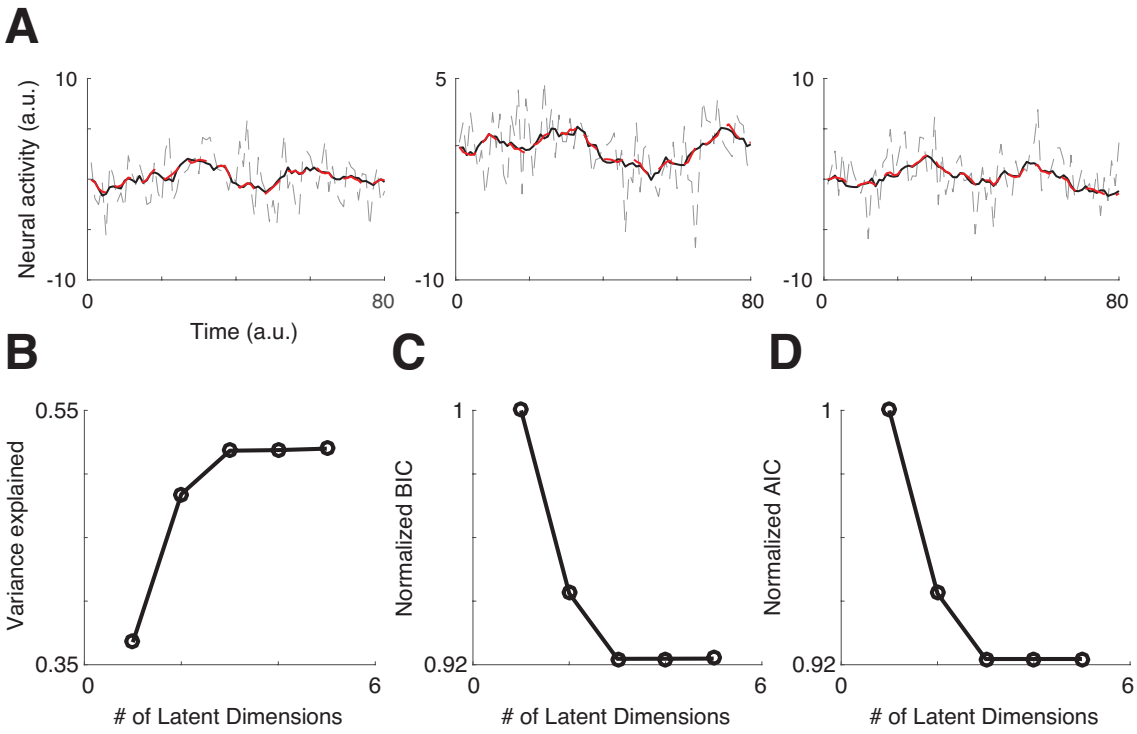


Figure A1.3: Performance of time-varying latent dynamical system analysis on the artificial data using Gaussian variability.
(**A**) Comparison between dynamics of artificial neural activity and its fit using LONO. Left to right panels exhibit activity of three randomly chosen artificial neurons (gray dash lines), their noiseless activity (black lines), and their estimated activity using LONO (red dash lines), in three single trials. Artificial neural activity was generated from **Equations A.1** and **A.2** with given parameter set Θ, where the dimension of latent space (**x**-space) is 3 and that of the neural space (**r**-space) is 10. Overall the estimation of the neural activity is close to that of noiseless version of the original neural dynamics. (**B**) The amount of explained variance in neural dynamics increases as a function of the latent dimension applied in TLDS fit, and it saturates when the effective number of the dimensions is close to the real latent dimension. (**C**) same as **B**, where the goodness-of-fit was measured using Bayesian information criterion (BIC, y-axis). (**D**) same as **B**, where the goodness-of-fit was measured using Akaike information criterion (AIC, y-axis).

with the estimation of $\mathbf{r}_t^k$ from the TLDS fit using the LONO method ($\hat{r}_{j,t}^l$, **Equation A.4**). In validating computer implementation of the fitting algorithm, we generated the dynamics of population activity (with n neurons and m neural modes), using a given parameter set Θ. We set a minimum number of constraints for parameter choices, where (1) matrices $\mathbf{R}$, $\mathbf{Q}$ and $\mathbf{Q}_0$ are diagonal matrices (for ease of simulation and traceability of the TLDS fit); (2) the diagonals of matrix $\mathbf{A} \in \mathbf{R}^{m \times m}$ are less than one; and (3) both matrices $\mathbf{A}$ and $\mathbf{C}$ ($\mathbf{C} \in \mathbf{R}^{n \times m}$) were random matrices with elements sampled from normal distribution $N(0, 1)$ (we made a post-hoc correction for collinearity across rows of matrices). For instance, we made an exemplary comparison based on a system with $n = 10$ neurons and $m = 3$ neural modes for 80 time steps. **Figure A1.3A** demonstrates the activity of three randomly picked exemplary neurons ($\mathbf{r}^k$; gray dash lines), and that of the noiseless version of the neural dynamics ($\bar{\mathbf{r}}^k$; black lines), and that of the LONO estimations from the other $n-1$ neurons ($\hat{\mathbf{r}}^k$; red dash lines), where the LONO estimation was based on the estimated parameter set Θ^* using m^*-dimensional TLDS system; $m^* = 3$. Notably, the activity of LONO estimations is close to that of the noiseless version of neural dynamics, which indicates a successful TLDS fit in the exemplary system.

The dimension of the latent space (the neural-mode dimension) is usually unknown and has to be estimated. The goodness-of-fit of the TLDS model should generally increase, until saturation, as a function of the latent dimension in the training dataset. Therefore, one main concern is the over-estimation of dimension of the neural modes,

where some neural modes would not contribute to the dynamics of the shared input or could overfit noise structure in the training dataset. Consequently, one needs to determine an optimal number of neural modes in TLDS fit based on their statistical significance.

As we discussed above, the amount of explained variance, R^2, based on LONO estimation could be one option for examining the minimum number of neural modes that explain most of the dynamics in the neural data. **Figure A1.3B** illustrates the increase of R^2 as a function of neural-modes and indicates a saturation of R^2 that started at 3. Of note, the saturation point in TLDS fit (namely, the effective number of neural modes) is equal to the real number of neural modes that generated the artificial neural data. One could thus potentially estimate the real number of neural modes to be the effective number from the dimension-R^2 curve. Importantly, in this case, the value of $R^2 = .53$ is very close to the maximum of $R^2_{\max} = .54$ based on the noiseless version of the neural data, $\bar{\mathbf{r}}^l_t$,

$$R^2_{\max} = 1 - \frac{\sum_{j=1}^{n} < r^l_{j,t} - \bar{r}^l_{j,t} >_{l,t}}{\sum_{j=1}^{n} < r^l_{j,t} - r_{j,0} >_{l,t}}$$

There are also other options for evaluating the effective number of latent modes, proposed in statistics based on the likelihood of the estimation (Gelman et al., 2013), such as the Bayesian information criterion (BIC; **Figure A1.3C**) (Schwarz and others, 1978) and the Akaike information criterion (AIC; **Figure A1.3D**) (Akaike, 1973).

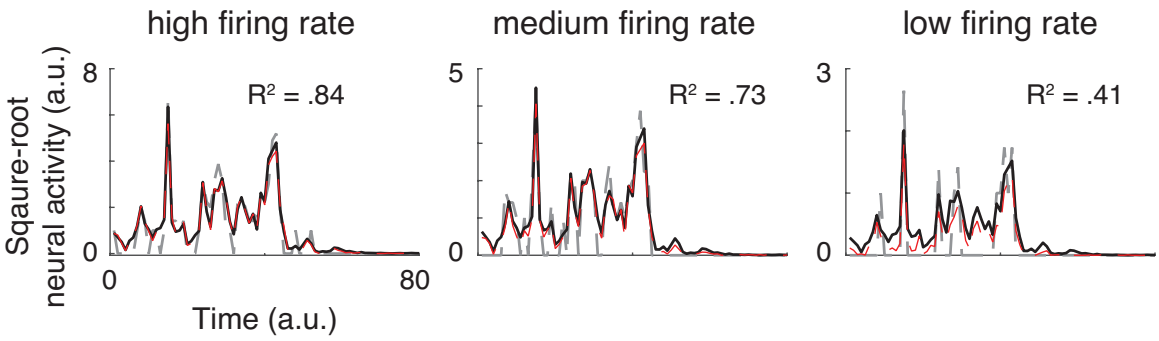


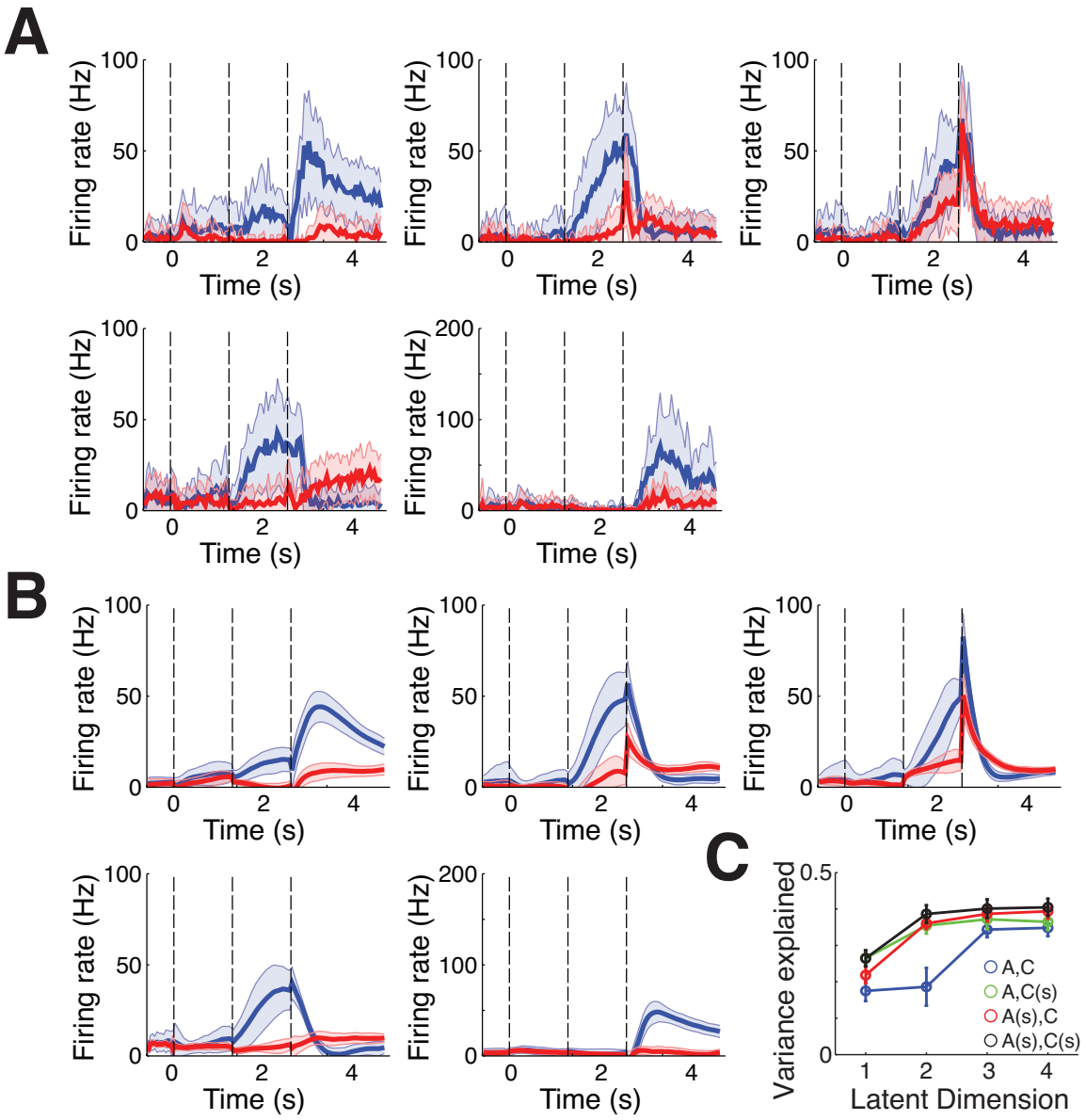
Figure A1.4: Performance of time-varying latent dynamical system analysis on the artificial data using Poisson variability.
(**A**) Comparison between dynamics of artificial neural activity and its fit using LONO. Left to right panels exhibit activity of the same artificial neurons (gray dash lines) which were scaled at different ratios (high ratio corresponding to a high rate of activity), and its noiseless activity at the corresponding firing-rate condition (black lines), and their estimated activity using LONO (red dash lines), in three single trials. Artificial neural activity was generated from **Equations A.2** and **A.5** with given parameter set Θ, where the dimension of latent space (**x**-space) is one and that of the neural space (**r**-space) is 10. Overall, the estimation of neural activity is close to that of the noiseless version of the original neural dynamics in the high rate condition. The estimation error increases as the firing rate get lower (shown as a decline of explained variance, R^2).

Both of these regularized the number of parameters in a fit explicitly upon the minimization of the negative log likelihood (Gelman et al., 2013). Nevertheless, in this example of artificial neural data, the effective number of latent modes was consistent across the different measurements of statistical significance.

Furthermore, we evaluated the performance of TLDS fit on the artificial data in a more realistic condition, where the generation of the spike count $\mathbf{r}_t^k$ follows Poisson variability in the "projection space". To simplify the comparison, we generated artificial neural data, where the dimension of latent modes is one, $m = 1$, and the dimension of neural space is 10, $n = 10$. Importantly, in a Poisson model (**Equation**

A.5), the variability is a linear function of the instantaneous mean firing rate; hence, the TLDS fit would be expected to fail, because it assumed the constant variability across time. To overcome this, we created a variance-stabilizing transform (square-root) of the spike counts in each time bin (where the variance converges to a constant at high firing rates) and applied the TLDS fit to the square-root of the neural data. Such a procedure is usually unnecessary for neurons that fire at high rates (data not shown). For this special case, we first compute the mean activity using **Equations A.2** and **A.6** for a given parameter set Θ. We then scale it at different ratios from 0.1, 0.5 to 1.0 to generate the low-, medium- and high firing rate conditions of the artificial neural data. As shown in the variance-stabilizing analysis, the Gaussian approximation of the Poisson model failed at low firing rate condition but improved as the firing rate increased. We therefore expect a good performance of TLDS fit in the high firing rate condition (**Figure A1.4**). To test this, we compared the real mean firing rate (modeled in **Equation A.5**) with the estimation of neural dynamics using LONO method after TLDS fit. The difference between them is minimal at high firing rate condition, but increases with lower firing rates. Accordingly, we also find the the same trends hold true for the explained variance, R^2. Of note, in the low firing rate condition, the LONO estimation tended to be lower than the real mean firing rate, which might signify the biased estimation of the rare events in a small sample size.

Lastly, we evaluated the performance of TLDS fit on the real simultaneous recordings. As discussed above, using artificial neural data based on the Poisson model, one


A
B
C
Firing rate (Hz)
Time (s)
Variance explained
Latent Dimension
A,C
A,C(s)
A(s),C
A(s),C(s)

would expected bad performance for the low firing neurons, due to the limitations of the TLDS fit. We therefore only applied the TLDS fit to the real simultaneous recordings using only neurons firing at high rates (**Figure A1.5A**). Unlike the test of Poisson model based on artificial data, we did not perform the variance stabilization procedure for the real neural data at high rates, because it led to little improvement (data not shown).

In this example (**Figure A1.5B**), we first fit the neural systems using a TLDS system with 3 latent neural modes and assumed that matrices **A** and **C** were epoch-dependent across trial. We then performed the LONO method to estimate the dynamics of a specific neuron based on the dynamics of the other four neurons on single trials. The estimation of the mean neural dynamics across each trial type showed significant similarity to that in real data (comparing dynamics of neurons in **Figures**

Figure A1.5 *(preceding page)*: Performance of time-varying latent dynamical system analysis on the real simultaneous neural recordings.
(**A-B**) Comparison between dynamics of simultaneous recorded neural activity and its fit using LONO. (**A**) The activity of five simultaneously recorded neurons in two trial type conditions ("trial type A", blue; "trial type B", red; **Figure 2.1A**), with four behavioral epochs (pre-sample, sample, delay and response, separated by vertical dash lines); solid line, mean activity; shaded area, sem. (**B**) same as **A** for LONO estimation of neural dynamics based on the 3-dimensional TLDS fit ($m = 3$, and $n = 5$). (**C**) Explained variance, R^2, in fits of different configurations of TLDS models; circle, mean performance for a 10-fold cross-validation; error bar, std.. Model 1 (blue), matrices **A** and **C** were constant across the task; Model 2 (green), matrix **A** was constant while matrix **C** was epoch-dependent across the task; Model 3 (red), matrix **C** was constant while matrix **A** was epoch-dependent across the task; Model 4 (black), matrices **A** and **C** were epoch-dependent across the task. Generally, the epoch-dependent models showed a better fit than the constant dynamical model (Model 1).

A1.5A and **A1.5B**), which indicates a good fit of the neural dynamics using the TLDS model.

To determine the effective number of latent neural modes and the effective complexity of the model, we performed cross-validation of the model with different configuration combinations. We ran trials with the number of latent neural modes from 1 to $n - 1$ (i.e. 4 in this example). We also tried tweaking the complexity of the model by constraining either matrix **A** or **C** to be a constant matrix across the task. In the latter case, the number of parameters to fit decreased when any matrices were constrained to be constant. In general, **Figure A1.5C** revealed that (1) the amount of explained variance increased with the number of latent neural modes and it saturated at 3; (2) the epoch-dependent implementation of the model is preferred, since the constant model (the blue line) shows a lower performance (χ^2 test, $p < .001$).

Bibliography

Abeles M (1982) Quantification, smoothing, and confidence limits for single-units' histograms. *Journal of Neuroscience Methods* 5:317–325.

Abeles M, Bergman H, Gat I, Meilijson I, Seidemann E, Tishby N, Vaadia E (1995) Cortical activity flips among quasi-stationary states. *Proceedings of the National Academy of Sciences of the United States of America* 92:8616–8620.

Afshar A, Santhanam G, Yu BM, Ryu SI, Sahani M, Shenoy KV (2011) Single-trial neural correlates of arm movement preparation. *Neuron* 71:555–564.

Ahrens MB, Engert F (2015) Large-scale imaging in small brains. *Current Opinion in Neurobiology* 32:78–86.

Ahrens MB, Li JM, Orger MB, Robson DN, Schier AF, Engert F, Portugues R (2012) Brain-wide neuronal dynamics during motor adaptation in zebrafish. *Nature* 485:471–477.

Ahrens MB, Orger MB, Robson DN, Li JM, Keller PJ (2013) Whole-brain func-

tional imaging at cellular resolution using light-sheet microscopy. *Nature Methods* 10:413–420.

Akaike H (1973) Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika* 60:255–265.

Akerboom J, Chen TW, Wardill TJ, Tian L, Marvin JS, Mutlu S, Calderon NC, Esposti F, Borghuis BG, Sun XR, Gordus A, Orger MB, Portugues R, Engert F, Macklin JJ, Filosa A, Aggarwal A, Kerr RA, Takagi R, Kracun S, Shigetomi E, Khakh BS, Baier H, Lagnado L, Wang SSH, Bargmann CI, Kimmel BE, Jayaraman V, Svoboda K, Kim DS, Schreiter ER, Looger LL (2012) Optimization of a GCaMP calcium indicator for neural activity imaging. *Journal of Neuroscience* 32:13819–13840.

Akhlaghpour H, Wiskerke J, Choi JY, Taliaferro JP, Au J, Witten I (2016) Dissociated sequential activity and stimulus encoding in the dorsomedial striatum during spatial working memory. *eLife* 5:e19507.

Albantakis L, Deco G (2011) Changes of mind in an attractor network of decision-making. *PLoS Computational Biology* 7:e1002086.

Aljadeff J, Lansdell BJ, Fairhall AL, Kleinfeld D (2016) Analysis of neuronal spike trains, deconstructed. *Neuron* 91:221–259.

Ames KC, Ryu SI, Shenoy KV (2014) Neural dynamics of reaching following incorrect or absent motor preparation. *Neuron* 81:438–451.

BIBLIOGRAPHY

Arieli A, Sterkin A, Grinvald A, Aertsen A (1996) Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science* 273:1868.

Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. *Nature Reviews Neuroscience* 7:358–366.

Barthelmé S, Mamassian P (2010) Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences of the United States of America* 107:20834–20839.

Bartho P, Curto C, Luczak A, Marguet SL, Harris KD (2009) Population coding of tone stimuli in auditory cortex: dynamic rate vector analysis. *European Journal of Neuroscience* 30:1767–1778.

Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J, Shadlen MN, Latham PE, Pouget A (2008) Probabilistic population codes for bayesian decision making. *Neuron* 60:1142–1152.

Berdondini L, Imfeld K, Maccione A, Tedesco M, Neukom S, Koudelka-Hep M, Martinoia S (2009) Active pixel sensor array for high spatio-temporal resolution electrophysiological recordings from single cell to large scale neuronal networks. *Lab on a Chip* 9:2644–2651.

Berger T, Borgdorff A, Crochet S, Neubauer FB, Lefort S, Fauvet B, Ferezou I, Carleton A, Lüscher HR, Petersen CC (2007) Combined voltage and calcium epi-

fluorescence imaging in vitro and in vivo reveals subthreshold and suprathreshold dynamics of mouse barrel cortex. *Journal of Neurophysiology* 97:3751–3762.

Bishop CM (1998) Latent variable models. In *Learning in Graphical Models*, pp. 371–403. Springer.

Bishop CM (2006) *Pattern recognition and Machine Learning.* Information Science and Statistics. Springer.

Bollimunta A, Totten D, Ditterich J (2012) Neural dynamics of choice: single-trial analysis of decision-related activity in parietal cortex. *Journal of Neuroscience* 32:12684–12701.

Boots B, Gordon G (2012) Two-manifold problems with applications to nonlinear system identification. In *International Conference on Machine Learning*.

Bradley A, Skottun BC, Ohzawa I, Sclar G, Freeman RD (1987) Visual orientation and spatial frequency discrimination: a comparison of single neurons and behavior. *Journal of Neurophysiology* 57:755–772.

Brendel W, Romo R, Machens CK (2011) Demixed principal component analysis. In *Advances in Neural Information Processing Systems*, pp. 2654–2662.

Brenner N, Bialek W, de Ruyter van Steveninck R (2000) Adaptive rescaling maximizes information transmission. *Neuron* 26:695–702.

BIBLIOGRAPHY

Briggman K, Abarbanel H, Kristan W (2005) Optical imaging of neuronal populations during decision-making. *Science* 307:896–901.

Brillinger DR (1992) Nerve cell spike train data analysis: a progression of technique. *Journal of the American Statistical Association* 87:260–271.

Britten KH, Shadlen MN, Newsome WT, Movshon JA (1993) Responses of neurons in macaque MT to stochastic motion signals. *Visual Neuroscience* 10:1157–1169.

Brody CD, Hernández A, Zainos A, Romo R (2003) Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cerebral Cortex* 13:1196–1207.

Broome BM, Jayaraman V, Laurent G (2006) Encoding and decoding of overlapping odor sequences. *Neuron* 51:467–482.

Brown EN, Kass RE, Mitra PP (2004) Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience* 7:456–461.

Brown SL, Joseph J, Stopfer M (2005) Encoding a temporally structured stimulus with a temporally structured neural representation. *Nature Neuroscience* 8:1568–1576.

Brunton BW, Botvinick MM, Brody CD (2013) Rats and humans can optimally accumulate evidence for decision-making. *Science* 340:95–98.

BIBLIOGRAPHY

Buesing L, Machado TA, Cunningham JP, Paninski L (2014) Clustered factor analysis of multineuronal spike data. In *Advances in Neural Information Processing Systems*, pp. 3500–3508.

Buesing L, Macke JH, Sahani M (2012) Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. In *Advances in Neural Information Processing Systems*, pp. 1682–1690.

Buzsaki G (2004) Large-scale recording of neuronal ensembles. *Nature Neuroscience* 7:446–451.

Buzsáki G (2015) Our skewed sense of space. *Science* 347:612–613.

Buzsáki G, Mizuseki K (2014) The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews Neuroscience* 15:264–278.

Carrillo-Reid L, Tecuapetla F, Tapia D, Hernández-Cruz A, Galarraga E, Drucker-Colin R, Bargas J (2008) Encoding network states by striatal cell assemblies. *Journal of Neurophysiology* 99:1435–1450.

Chen TW, Wardill TJ, Sun Y, Pulver SR, Renninger SL, Baohan A, Schreiter ER, Kerr RA, Orger MB, Jayaraman V, Looger LL, Svoboda K, Kim DS (2013) Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* 499:295–300.

Cheung KC (2007) Implantable microscale neural interfaces. *Biomedical Microdevices* 9:923–938.

BIBLIOGRAPHY

Chornoboy E, Schramm L, Karr A (1988) Maximum likelihood identification of neural point process systems. *Biological Cybernetics* 59:265–275.

Churchland AK, Kiani R, Chaudhuri R, Wang XJ, Pouget A, Shadlen MN (2011) Variance as a signature of neural computations during decision making. *Neuron* 69:818–831.

Churchland AK, Kiani R, Shadlen MN (2008) Decision-making with multiple alternatives. *Nature Neuroscience* 11:693–702.

Churchland MM, Cunningham JP, Kaufman MT, Foster JD, Nuyujukian P, Ryu SI, Shenoy KV (2012) Neural population dynamics during reaching. *Nature* 487:1–23.

Churchland MM, Cunningham JP, Kaufman MT, Ryu SI, Shenoy KV (2010a) Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron* 68:387–400.

Churchland MM, Yu BM, Cunningham JP, Sugrue LP, Cohen MR, Corrado GS, Newsome WT, Clark AM, Hosseini P, Scott BB, Bradley DC, Smith MA, Kohn A, Movshon JA, Armstrong KM, Moore T, Chang SW, Snyder LH, Lisberger SG, Priebe NJ, Finn IM, Ferster D, Ryu SI, Santhanam G, Sahani M, Shenoy KV (2010b) Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience* 13:369–378.

Cohen MR, Maunsell JHR (2010) A neuronal population measure of atten-

tion predicts behavioral performance on individual trials. *Journal of Neuroscience* 30:15241–15253.

Cohen MR, Maunsell JHR (2009) Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience* 12:1594–1600.

Cohen MR, Newsome WT (2008) Context-dependent changes in functional circuitry in visual area MT. *Neuron* 60:162–173.

Compte A, Brunel N, Goldman-Rakic PS, Wang XJ (2000) Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex* 10:910–923.

Cook EP, Maunsell JHR (2002) Dynamics of neuronal responses in macaque MT and VIP during motion detection. *Nature Neuroscience* 5:985–994.

Coyer TM (2014) Learning in pattern recognition. In *Methodologies of Pattern Recognition*, p. 111. Academic Press.

Cunningham JP, Yu BM (2014) Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience* 17:1500–1509.

Curtis CE, Lee D (2010) Beyond working memory: the role of persistent activity in decision making. *Trends in Cognitive Sciences* 14:216–222.

Dana H, Chen TW, Hu A, Shields BC, Guo C, Looger LL, Kim DS, Svoboda K

BIBLIOGRAPHY

(2014) Thy1-GCaMP6 transgenic mice for neuronal population imaging in vivo. *PLoS One* 9:e108697–9.

Dana H, Mohar B, Sun Y, Narayan S, Gordus A, Hasseman JP, Tsegaye G, Holt GT, Hu A, Walpita D, Patel R, Macklin JJ, Bargmann CI, Ahrens MB, Schreiter ER, Jayaraman V, Looger LL, Svoboda K, Kim DS (2016) Sensitive red protein calcium indicators for imaging neural activity. *eLife* 5:413.

Danóczy M, Hahnloser R (2006) Efficient estimation of hidden state dynamics from spike trains. In *Advances in Neural Information Processing Systems*, pp. 227–234.

Dayan P, Abbott LF (2001) *Theoretical neuroscience.* Cambridge, MA: MIT Press.

Dean A (1981) The variability of discharge of simple cells in the cat striate cortex. *Experimental Brain Research* 44:437–440.

Del Cul A, Dehaene S, Reyes P, Bravo E, Slachevsky A (2009) Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain* 132:2531–2540.

Dienes Z, Seth A (2010) Gambling on the unconscious: a comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition* 19:674–681.

Dombeck DA, Khabbaz AN, Collman F, Adelman TL, Tank DW (2007) Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* 56:43–57.

BIBLIOGRAPHY

Druckmann S, Chklovskii DB (2010) Over-complete representations on recurrent neural networks can support persistent percepts. In *Advances in Neural Information Processing Systems*, pp. 541–549.

Druckmann S, Chklovskii DB (2012) Neuronal circuits underlying persistent representations despite time varying activity. *Current Biology* 22:2095–2103.

Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A (2012) The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience* 32:3612–3628.

Drugowitsch J, Moreno-Bote R, Pouget A (2014) Relation between belief and performance in perceptual decision making. *PLoS One* 9:e96511–16.

Durstewitz D, Vittoz NM, Floresco SB, Seamans JK (2010) Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron* 66:438–448.

Engel TA, Chaisangmongkon W, Freedman DJ, Wang XJ (2015) Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nature Communications* 6:1–12.

Everitt BS (1984) *An introduction to latent variable models.* Springer Science & Business Media.

Fejtl M, Stett A, Nisch W, Boven KH, Möller A (2006) On micro-electrode array

revival: its development, sophistication of recording, and stimulation. In *Advances in Network Electrophysiology*, pp. 24–37. Springer.

Fetsch CR, Kiani R, Newsome WT, Shadlen MN (2014) Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* 83:797 – 804.

Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7:179–188.

Flavell JH (1979) Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* 34:906–911.

Fleming SM, Dolan RJ (2010) Effects of loss aversion on post-decision wagering: implications for measures of awareness. *Consciousness and Cognition* 19:352–363.

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. *Science* 329:1541–1543.

Freeman J (2015) Open source tools for large-scale neuroscience. *Current Opinion in Neurobiology* 32:156–163.

Freeman J, Vladimirov N, Kawashima T, Mu Y, Sofroniew NJ, Bennett DV, Rosen J, Yang CT, Looger LL, Ahrens MB (2014) Mapping brain activity at scale with cluster computing. *Nature Methods* 11:941–950.

Frey U, Egert U, Heer F, Hafizovic S, Hierlemann A (2009) Microelectronic system

for high-resolution mapping of extracellular electric fields applied to brain slices. *Biosensors and Bioelectronics* 24:2191–2198.

Friedman J, Hastie T, Tibshirani R (2001) *The elements of statistical learning.* Springer series in statistics, Springer, Berlin.

Fromherz P (2006) Three levels of neuroelectronic interfacing. *Annals of the New York Academy of Sciences* 1093:143–160.

Fu Y, Tucciarone JM, Espinosa JS, Sheng N, Darcy DP, Nicoll RA, Huang ZJ, Stryker MP (2014) A cortical circuit for gain control by behavioral state. *Cell* 156:1139–1152.

Fukunaga R (1990) *Introduction to statistical pattern recognition.* Academic Press.

Furman M, Wang XJ (2008) Similarity effect and optimal control of multiple-choice decision making. *Neuron* 60:1153–1168.

Galvani L, Aldini G (1792) *De viribus electricitatis in motu musculari comentarius cum joannis aldini dissertatione et notis; accesserunt epistolae ad animalis electricitatis theoriam pertinentes.* Apud Societatem Typographicam.

Ganguly K, Carmena JM (2009) Emergence of a stable cortical map for neuroprosthetic control. *PLoS Biology* 7:1–13.

Ganmor E, Krumin M, Rossi LF, Carandini M, Simoncelli EP (2016) Direct estimation of firing rates from calcium imaging data. *arXiv:1601.00364* .

BIBLIOGRAPHY

Gao Y, Busing L, Shenoy KV, Cunningham JP (2015) High-dimensional neural spike train analysis with generalized count linear dynamical systems In *Advances in Neural Information Processing Systems*, pp. 2044–2052.

Geisler WS, Albrecht DG (1997) Visual cortex neurons in monkeys and cats: detection, discrimination, and identification. *Visual Neuroscience* 14:897–919.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) *Bayesian data analysis.* CRC Press.

Gerstein GL, Perkel DH (1969) Simultaneously recorded trains of action potentials: analysis and functional interpretation. *Science* 164:828–830.

Ghahramani Z, Hinton GE (1996a) Parameter estimation for linear dynamical systems. Technical report, Department of Computer Science, University of Toronto.

Ghahramani Z, Hinton GE (1996b) Switching state-space models. Technical report, Department of Computer Science, University of Toronto.

Gilja V, Nuyujukian P, Chestek CA, Cunningham JP, Yu BM, Fan JM, Churchland MM, Kaufman MT, Kao JC, Ryu SI, Shenoy KV (2012) A high-performance neural prosthesis enabled by control algorithm design. *Nature Neuroscience* 15:1752–1757.

Gold JI, Shadlen MN (2007) The neural basis of decision making. *Annual Review of Neuroscience* 30:535–574.

BIBLIOGRAPHY

Graziano M, Parra LC, Sigman M (2010) Neurophysiology of perceived confidence. In *Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 2818–2821.

Graziano M, Parra LC, Sigman M (2015) Neural Correlates of Perceived Confidence in a Partial Report Paradigm. *Journal of Cognitive Neuroscience* 27:1090–1103.

Graziano M, Sigman M (2009) The spatial and temporal construction of confidence in the visual scene. *PLoS One* 4:e4909.

Greenberg DS, Houweling AR, Kerr JN (2008) Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nature Neuroscience* 11:749–751.

Grewe BF, Langer D, Kasper H, Kampa BM, Helmchen F (2010) High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. *Nature Methods* 7:399–405.

Grienberger C, Konnerth A (2012) Imaging calcium in neurons. *Neuron* 73:862–885.

Guo JZ, Graves AR, Guo WW, Zheng J, Lee A, Rodríguez-González J, Li N, Macklin JJ, Phillips JW, Mensh BD, Branson K, Hantman AW (2015) Cortex commands the performance of skilled movement. *eLife* 4:219.

Guo Y, Hastie T, Tibshirani R (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* 8:86–100.

BIBLIOGRAPHY

Guo ZV, Hires SA, Li N, O'Connor DH, Komiyama T, Ophir E, Huber D, Bonardi C, Morandell K, Gutnisky D, Peron S, Xu Nl, Cox J, Svoboda K (2014a) Procedures for behavioral experiments in head-fixed mice. *PLoS One* 9:1–16.

Guo ZV, Li N, Huber D, Ophir E, Gutnisky D, Ting JT, Feng G, Svoboda K (2014b) Flow of cortical activity underlying a tactile decision in mice. *Neuron* 81:179–194.

Hanks TD, Ditterich J, Shadlen MN (2006) Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience* 9:682–689.

Hansel D, Mato G, Meunier C, Neltner L (1998) On numerical simulations of integrate-and-fire neural networks. *Neural Computation* 10:467–483.

Harris KD, Quiroga RQ, Freeman J, Smith SL (2016) Improving data quality in neuronal population recordings. *Nature Neuroscience* 19:1165–1174.

Harvey CD, Coen P, Tank DW (2012) Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* 484:62–68.

Heeger DJ (1992) Normalization of cell responses in cat striate cortex. *Visual Neuroscience* 9:181–197.

Hell SW (2007) Far-field optical nanoscopy. *Science* 316:1153–1158.

Hell SW (2010) Far-field optical nanoscopy In *Single Molecule Spectroscopy in Chemistry, Physics and Biology*, pp. 365–398. Springer.

BIBLIOGRAPHY

Heller I, Kong J, Heering HA, Williams KA, Lemay SG, Dekker C (2005) Individual single-walled carbon nanotubes as nanoelectrodes for electrochemistry. *Nano Letters* 5:137–142.

Hochberg LR, Serruya MD, Friehs GM, Mukand JA, Saleh M, Caplan AH, Branner A, Chen D, Penn RD, Donoghue JP (2006) Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* 442:164–171.

Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55–67.

Hromádka T, DeWeese MR, Zador AM (2008) Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biology* 6:e16.

Huber D, Gutnisky DA, Peron S, O'Connor DH, Wiegert JS, Tian L, Oertner TG, Looger LL, Svoboda K (2012) Multiple dynamic representations in the motor cortex during sensorimotor learning. *Nature* 484:473–478.

Hutzler M, Lambacher A, Eversmann B, Jenkner M, Thewes R, Fromherz P (2006) High-resolution multitransistor array recording of electrical field potentials in cultured brain slices. *Journal of Neurophysiology* 96:1638–1645.

Huys R, Braeken D, Jans D, Stassen A, Collaert N, Wouters J, Loo J, Severi S, Vleugels F, Callewaert G et al. (2012) Single-cell recording and stimulation with

a 16k micro-nail electrode array integrated on a 0.18 μm cmos chip. *Lab on a Chip* 12:1274–1280.

Inoue M, Takeuchi A, Horigane Si, Ohkura M, Gengyo-Ando K, Fujii H, Kamijo S, Takemoto-Kimura S, Kano M, Nakai J et al. (2015) Rational design of a high-affinity, fast, red calcium indicator R-CaMP2. *Nature Methods* 12:64–70.

Insabato A, Pannunzi M, Rolls ET, Deco G (2010) Confidence-related decision-making. *Journal of Neurophysiology* 104:539–547.

Ji N, Freeman J, Smith SL (2016) Technologies for imaging neural activity in large volumes. *Nature Neuroscience* 19:1154–1164.

Ji N, Shroff H, Zhong H, Betzig E (2008) Advances in the speed and resolution of light microscopy. *Current Opinion in Neurobiology* 18:605–616.

Jolliffe I (2014) *Principal component analysis.* Wiley StatsRef: Statistics Reference Online.

Jones LM, Fontanini A, Sadacca BF, Miller P, Katz DB (2007) Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proceedings of the National Academy of Sciences of the United States of America* 104:18772–18777.

Juslin P, Olsson H (1997) Thurstonian and Brunswikian origins of uncertainty in judgment: a sampling model of confidence in sensory discrimination. *Psychological Review* 104:344–366.

BIBLIOGRAPHY

Juslin P, Olsson N (1996) Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology Learning Memory and Cognition* 22:1304–1316.

Jvreskog K (1996) *Applied factor analysis in the natural sciences.* Cambridge University Press.

Kao JC, Nuyujukian P, Ryu SI, Churchland MM, Cunningham JP, Shenoy KV (2015) Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. *Nature Communications* 6:7759.

Kass RE, Ventura V, Cai C (2003) Statistical smoothing of neuronal data. *Network-Computation in Neural Systems* 14:5–16.

Katzner S, Nauhaus I, Benucci A, Bonin V, Ringach DL, Carandini M (2009) Local origin of field potentials in visual cortex. *Neuron* 61:35–41.

Kaufman MT, Churchland MM, Ryu SI, Shenoy KV (2014) Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience* 17:440–448.

Kaufman MT, Churchland MM, Ryu SI, Shenoy KV (2015) Vacillation, indecision and hesitation in moment-by-moment decoding of monkey motor cortex. *eLife* 4:e04677.

BIBLIOGRAPHY

Keller PJ, Ahrens MB (2015) Visualizing whole-brain activity and development at the single-cell level using light-sheet microscopy. *Neuron* 85:462–483.

Kemere C, Santhanam G, Byron MY, Afshar A, Ryu SI, Meng TH, Shenoy KV (2008) Detecting neural-state transitions using hidden markov models for motor cortical prostheses. *Journal of Neurophysiology* 100:2441–2452.

Kepecs A, Mainen ZF (2012) A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367:1322–1337.

Kepecs A, Uchida N, Zariwala HA, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455:227–231.

Kiani R, Shadlen MN (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* 324:759–764.

Kiani R, Corthell L, Shadlen MN (2014) Choice certainty is informed by both evidence and decision time. *Neuron* 84:1329–1342.

Kiani R, Cueva CJ, Reppas JB, Peixoto D, Ryu SI, Newsome WT (2015) Natural grouping of neural responses reveals spatially segregated clusters in prearcuate cortex. *Neuron* 85:1–35.

Kiani R, Hanks TD, Shadlen MN (2008) Bounded integration in parietal cortex un-

derlies decisions even when viewing duration is dictated by the environment. *Journal of Neuroscience* 28:3017–3029.

Knott M, Bartholomew DJ (1999) *Latent variable models and factor analysis.* Edward Arnold.

Kobak D, Brendel W, Constantinidis C, Feierstein CE (2016) Demixed principal component analysis of neural population data. *eLife* 5:e10989.

Komiyama T, Sato TR, O'Connor DH, Zhang YX, Huber D, Hooks BM, Gabitto M, Svoboda K (2010) Learning-related fine-scale specificity imaged in motor cortex circuits of behaving mice. *Nature* 464:1182–1186.

Komura Y, Nikkuni A, Hirashima N, Uetake T, Miyamoto A (2013) Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nature Neuroscience* 16:749–755.

Krapf D, Wu MY, Smeets RM, Zandbergen HW, Dekker C, Lemay SG (2006) Fabrication and characterization of nanopore-based electrodes with radii down to 2 nm. *Nano Letters* 6:105–109.

Krzanowski W (2000) *Principles of multivariate analysis.* OUP Oxford.

Kulkarni JE, Paninski L (2007) Common-input models for multiple neural spike-train data. *Network: Computation in Neural Systems* 18:375–407.

BIBLIOGRAPHY

Lak A, Costa GM, Romberg E, Koulakov AA, Mainen ZF, Kepecs A (2014) Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* 84:190–201.

Latimer KW, Yates JL, Meister MLR, Huk AC, Pillow JW (2015) Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* 349:184–187.

Lawhern V, Wu W, Hatsopoulos N, Paninski L (2010) Population decoding of motor cortical activity using a generalized linear model with hidden states. *Journal of Neuroscience Methods* 189:267–280.

Li N, Chen TW, Guo ZV, Gerfen CR, Svoboda K (2015) A motor cortex circuit for motor planning and movement. *Nature* 519:1–16.

Li N, Daie K, Svoboda K, Druckmann S (2016) Robust neuronal dynamics in premotor cortex during motor planning. *Nature* 532:1–25.

Lim S, McKee JL, Woloszyn L, Amit Y, Freedman DJ, Sheinberg DL, Brunel N (2015) Inferring learning rules from distributions of firing rates in cortical neurons. *Nature Neuroscience* 18:1804–1810.

Liu F, Wang XJ (2008) A common cortical circuit mechanism for perceptual categorical discrimination and veridical judgment. *PLoS Computational Biology* 4:e1000253.

BIBLIOGRAPHY

Ljung L (1998) System identification. In *Signal Analysis and Prediction*, pp. 163–173. Springer.

Luo L, Callaway EM, Svoboda K (2008) Genetic dissection of neural circuits. *Neuron* 57:634–660.

Machens CK (2010) Demixing population activity in higher cortical areas. *Frontiers in Computational Neuroscience* 4:126.

Machens CK, Romo R, Brody CD (2010) Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *Journal of Neuroscience* 30:350–360.

Macke JH, Buesing L, Cunningham JP, Yu BM, Shenoy KV, Sahani M (2011) Empirical models of spiking in neural populations In *Advances in Neural Information Processing Systems*, pp. 1350–1358.

Malvache A, Reichinnek S, Villette V, Haimerl C, Cossart R (2016) Awake hippocampal reactivations project onto orthogonal neuronal assemblies. *Science* 353:1280–1283.

Mante V, Sussillo D, Shenoy KV, Newsome WT (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503:1–19.

Maravall M, Mainen ZF, Sabatini BL, Svoboda K (2000) Estimating intracellular calcium concentrations and buffering without wavelength ratioing. *Biophysical Journal* 78:2655–2667.

BIBLIOGRAPHY

Matsuzaki M, Ellis-Davies GC, Nemoto T, Miyashita Y, Iino M, Kasai H (2001) Dendritic spine geometry is critical for AMPA receptor expression in hippocampal CA1 pyramidal neurons. *Nature Neuroscience* 4:1086–1092.

Mazor O, Laurent G (2005) Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron* 48:661–673.

Mazurek ME, Roitman JD, Ditterich J, Shadlen MN (2003) A role for neural integrators in perceptual decision making. *Cerebral Cortex* 13:1257–1269.

McCullagh P, Nelder JA (1989) *Generalized linear models.* CRC press.

Middlebrooks PG, Sommer MA (2011) Metacognition in monkeys during an oculomotor task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37:325–337.

Middlebrooks PG, Sommer MA (2012) Neuronal correlates of metacognition in primate frontal cortex. *Neuron* 75:517–530.

Miller EK, Wilson MA (2008) All my circuits: using multiple electrodes to understand functioning neural networks. *Neuron* 60:483–488.

Mizrahi A, Crowley JC, Shtoyerman E, Katz LC (2004) High-resolution in vivo imaging of hippocampal dendrites and spines. *Journal of Neuroscience* 24:3147–3151.

Mizuseki K, Buzsáki G (2013) Preconfigured, skewed distribution of firing rates in the hippocampus and entorhinal cortex. *Cell Reports* 4:1010–1021.

BIBLIOGRAPHY

Morcos AS, Harvey CD (2016) History-dependent variability in population dynamics during evidence accumulation in cortex. *Nature Neuroscience* 19:1672–1681.

Moreno-Bote R (2010) Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. *Neural Computation* 22:1786–1811.

Nam Y, Wheeler BC (2011) In vitro microelectrode array technology and neural recordings. *Critical Reviews in Biomedical Engineering* 39:45–61.

Nicolelis MA (2007) *Methods for neural ensemble recordings.* CRC press.

Nicolelis MA, Dimitrov D, Carmena JM, Crist R, Lehew G, Kralik JD, Wise SP (2003) Chronic, multisite, multielectrode recordings in macaque monkeys. *Proceedings of the National Academy of Sciences of the United States of America* 100:11041–11046.

Nimchinsky EA, Sabatini BL, Svoboda K (2002) Structure and function of dendritic spines. *Annual Review of Physiology* 64:313–353.

O'Connor DH, Hires SA, Guo ZV, Li N, Yu J, Sun QQ, Huber D, Svoboda K (2013) Neural coding during active somatosensation revealed using illusory touch. *Nature Neuroscience* 16:958–965.

O'Connor DH, Peron SP, Huber D, Svoboda K (2010) Neural activity in barrel cortex underlying vibrissa-based object localization in mice. *Neuron* 67:1048–1061.

Ohki K, Chung S, Ch'ng YH, Kara P, Reid RC (2005) Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* 433:597–603.

Ohkura M, Sasaki T, Sadakari J, Gengyo-Ando K, Kagawa-Nagamura Y, Kobayashi C, Ikegaya Y, Nakai J (2012) Genetically encoded green fluorescent Ca^{2+} indicators with improved detectability for neuronal Ca^{2+} signals. *PLoS One* 7:e51286.

Oñativia J, Schultz SR, Dragotti PL (2013) A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging. *Journal of Neural Engineering* 10:046017–15.

Palmer LM, Stuart GJ (2009) Membrane potential changes in dendritic spines during action potentials and synaptic input. *Journal of Neuroscience* 29:6897–6903.

Paninski L (2004) Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems* 15:243–262.

Paninski L, Ahmadian Y, Ferreira DG, Koyama S, Rad KR, Vidne M, Vogelstein J, Wu W (2010) A new look at state-space models for neural data. *Journal of Computational Neuroscience* 29:107–126.

Park IM, Meister MLR, Huk AC, Pillow JW (2014) Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nature Neuroscience* 17:1395–1403.

Park IJ, Bobkov YV, Ache BW, Principe JC (2013) Quantifying bursting neuron

activity from calcium signals using blind deconvolution. *Journal of Neuroscience Methods* 218:196–205.

Peron S, Chen TW, Svoboda K (2015) Comprehensive imaging of cortical networks. *Current Opinion in Neurobiology* 32:115–123.

Peron S, Svoboda K (2010) From cudgel to scalpel: toward precise neural control with optogenetics. *Nature Methods* 8:30–34.

Peron SP, Freeman J, Iyer V, Guo C, Svoboda K (2015) A Cellular Resolution Map of Barrel Cortex Activity during Tactile Behavior. *Neuron* 86:783–799.

Persaud N, Mcleod P, Cowey A (2007) Post-decision wagering objectively measures awareness. *Nature Neuroscience* 10:257–261.

Peters AJ, Chen SX, Komiyama T (2014) Emergence of reproducible spatiotemporal activity during motor learning. *Nature* 510:263–267.

Petreska B, Byron MY, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M (2011) Dynamical segmentation of single trials from population neural data. In *Advances in Neural Information Processing Systems*, pp. 756–764.

Pfau D, Pnevmatikakis EA, Paninski L (2013) Robust learning of low-dimensional dynamics from large neural ensembles. In *Advances in Neural Information Processing Systems*, pp. 2391–2399.

BIBLIOGRAPHY

Picardo MA, Merel J, Katlowitz KA, Vallentin D, Okobi DE, Benezra SE, Clary RC, Pnevmatikakis EA, Paninski L, Long MA (2016) Population-level representation of a temporal sequence underlying song production in the zebra finch. *Neuron* 90:866–876.

Pierrel R, Murray CS (1963) Some relationships between comparative judgment, confidence, and decision-time in weight-lifting. *American Journal of Psychology* 76:28–38.

Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, Simoncelli EP (2008) Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* 454:995–999.

Pine J (2006) A history of MEA development. In *Advances in Network Electrophysiology*, pp. 3–23. Springer.

Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400:233–238.

Pleskac TJ, Busemeyer JR (2010) Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review* 117:864–901.

Pnevmatikakis EA, Merel J, Pakman A, Paninski L (2014a) Bayesian spike inference from calcium imaging data. In *Asilomar Conference on Signals, Systems, and Computers*.

Pnevmatikakis EA, Gao Y, Soudry D, Pfau D, Lacefield C, Poskanzer K, Bruno R, Yuste R, Paninski L (2014b) A structured matrix factorization framework for large scale calcium imaging data analysis. *arXiv:1409.2903* .

Pnevmatikakis EA, Soudry D, Gao Y, Machado TA, Merel J, Pfau D, Reardon T, Mu Y, Lacefield C, Yang W, Ahrens M, Bruno R, Jessell TM, Peterka DS, Yuste R, Paninski L (2016) Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron* 89:285–299.

Pologruto TA, Yasuda R, Svoboda K (2004) Monitoring neural activity and [Ca^{2+}] with genetically encoded Ca^{2+} indicators. *Journal of Neuroscience* 24:9572–9579.

Ponce-Alvarez A, Nácher V, Luna R, Riehle A, Romo R (2012) Dynamics of cortical neuronal ensembles transit from decision making to storage for later report. *Journal of Neuroscience* 32:11956–11969.

Qiao Y, Chen J, Guo X, Cantrell D, Ruoff R, Troy J (2005) Fabrication of nanoelectrodes for neurophysiology: cathodic electrophoretic paint insulation and focused ion beam milling. *Nanotechnology* 16:1598.

Rabiner L, Juang B (1986) An introduction to hidden markov models. *IEEE ASSP Magazine* 3:4–16.

Rao CR (1948) The utilization of multiple measurements in problems of biologi-

cal classification. *Journal of the Royal Statistical Society. Series B (Methodological)* 10:159–203.

Ratcliff R, Smith PL (2004) A comparison of sequential sampling models for two-choice reaction time. *Psychological Review* 111:333–367.

Ratcliff R, Starns JJ (2009) Modeling confidence and response time in recognition memory. *Psychological Review* 116:59–83.

Resulaj A, Kiani R, Wolpert DM, Shadlen MN (2009) Changes of mind in decision-making. *Nature* 461:263–266.

Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. *Nature* 497:1–58.

Roitman JD, Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience* 22:9475–9489.

Rolls ET, Grabenhorst F, Deco G (2010a) Choice, difficulty, and confidence in the brain. *NeuroImage* 53:694–706.

Rolls ET, Grabenhorst F, Deco G (2010b) Decision-making, errors, and confidence in the brain. *Journal of Neurophysiology* 104:2359–2374.

Romo R, Brody CD, Hernández A, Lemus L (1999) Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* 399:470–473.

BIBLIOGRAPHY

Roweis S, Ghahramani Z (1999) A unifying review of linear Gaussian models. *Neural Computation* 11:305–345.

Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326.

Rupasov VI, Lebedev MA, Erlichman JS, Linderman M (2012) Neuronal variability during handwriting: lognormal distribution. *PLoS One* 7:e34759.

Sabatini BL, Oertner TG, Svoboda K (2002) The life cycle of Ca^{2+} ions in dendritic spines. *Neuron* 33:439–452.

Sabatini BL, Svoboda K (2000) Analysis of calcium channels in single spines using optical fluctuation analysis. *Nature* 408:589–593.

Sadtler PT, Quick KM, Golub MD, Chase SM, Ryu SI, Tyler-Kabara EC, Byron MY, Batista AP (2014) Neural constraints on learning. *Nature* 512:423–426.

Saha D, Leong K, Li C, Peterson S, Siegel G, Raman B (2013) A spatiotemporal coding mechanism for background-invariant odor recognition. *Nature Neuroscience* 16:1830–1839.

Santhanam G, Byron MY, Gilja V, Ryu SI, Afshar A, Sahani M, Shenoy KV (2009) Factor-analysis methods for higher-performance neural prostheses. *Journal of Neurophysiology* 102:1315–1330.

BIBLIOGRAPHY

Sasaki T, Takahashi N, Matsuki N, Ikegaya Y (2008) Fast and accurate detection of action potentials from somatic calcium fluctuations. *Journal of Neurophysiology* 100:1668–1676.

Scanziani M, Hausser M (2009) Electrophysiology in the age of light. *Nature* 461:930–939.

Scheuss V, Yasuda R, Sobczyk A, Svoboda K (2006) Nonlinear [Ca^{2+}] signaling in dendrites and spines caused by activity-dependent depression of Ca^{2+} extrusion. *Journal of Neuroscience* 26:8183–8194.

Schwartz AB (2004) Cortical neural prosthetics. *Annual Review of Neuroscience* 27:487–507.

Schwarz G et al. (1978) Estimating the dimension of a model. *The Annals of Statistics* 6:461–464.

Scobey R, Gabor A (1989) Orientation discrimination sensitivity of single units in cat primary visual cortex. *Experimental Brain Research* 77:398–406.

Seber GA (2009) *Multivariate observations.* John Wiley & Sons.

Seidemann E, Meilijson I, Abeles M, Bergman H, Vaadia E (1996) Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task. *Journal of Neuroscience* 16:752–768.

BIBLIOGRAPHY

Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology* 86:1916–1936.

Shadlen MN, Kiani R (2013) Decision making as a window on cognition. *Neuron* 80:791–806.

Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of Neuroscience* 18:3870–3896.

Shadlen MN, Shohamy D (2016) Decision making and sequential sampling from memory. *Neuron* 90:927–939.

Shenoy KV, Kaufman MT, Sahani M, Churchland MM (2011) A dynamical systems view of motor preparation: implications for neural prosthetic system design. *Progress in Brain Research* 192:33–58.

Shibata R (1989) Statistical aspects of model selection. In *From data to model*, pp. 215–240. Springer.

Shumway RH, Stoffer DS (1982) An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* 3:253–264.

Smith AC, Brown EN (2003) Estimating a state-space model from point process observations. *Neural Computation* 15:965–991.

Smith JD (2009) The study of animal metacognition. *Trends in Cognitive Sciences* 13:389–396.

Smith JD, Shields WE, Washburn DA (2003) The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences* 26:317–39.

Smith SL, Judy JW, Otis TS (2004) An ultra small array of electrodes for stimulating multiple inputs into a single neuron. *Journal of Neuroscience Methods* 133:109–114.

Snowden RJ, Treue S, Andersen RA (1992) The response of neurons in areas V1 and MT of the alert rhesus monkey to moving random dot patterns. *Experimental Brain Research* 88:389–400.

Sofroniew NJ, Flickinger D, King J, Svoboda K (2016) A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *eLife* 5:413.

Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *Journal of Neuroscience* 13:334–350.

Soltani A, Wang XJ (2006) A biophysically based neural model of matching law behavior: melioration by stochastic synapses. *Journal of Neuroscience* 26:3731–3744.

Spira ME, Hai A (2013) Multi-electrode array technologies for neuroscience and cardiology. *Nature Nanotechnology* 8:83–94.

Stevenson IH, Kording KP (2011) How advances in neural recording affect data analysis. *Nature Neuroscience* 14:139–142.

BIBLIOGRAPHY

Stone M (1977) An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 44–47.

Stopfer M, Jayaraman V, Laurent G (2003) Intensity versus identity coding in an olfactory system. *Neuron* 39:991–1004.

Stosiek C, Garaschuk O, Holthoff K, Konnerth A (2003) In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences of the United States of America* 100:7319–7324.

Stuphorn V, Brown JW, Schall JD (2010) Role of supplementary eye field in saccade initiation: executive, not direct, control. *Journal of Neurophysiology* 103:801–816.

Stuphorn V, Schall JD (2006) Executive control of countermanding saccades by the supplementary eye field. *Nature Neuroscience* 9:925–931.

Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. *Science* 304:1782–1787.

Sun W, Tan Z, Mensh BD, Ji N (2015) Thalamus provides layer 4 of primary visual cortex with orientation-and direction-tuned inputs. *Nature Neuroscience* 19:308–315.

Sussillo D, Churchland MM, Kaufman MT, Shenoy KV (2015) A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience* 18:1025–1033.

BIBLIOGRAPHY

Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319–2323.

Theis L, Berens P, Froudarakis E, Reimer J, Roman-Roson M, Baden T, Euler T, Tolias AS, Bethge M (2016) Benchmarking spike rate inference in population calcium imaging. *Neuron* 90:471–482.

Tian L, Hires SA, Mao T, Huber D, Chiappe ME, Chalasani SH, Petreanu L, Akerboom J, McKinney SA, Schreiter ER, Bargmann CI, Jayaraman V, Svoboda K, Looger LL (2009) Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nature Methods* 6:875–881.

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 99:6567–6572.

Tobler PN, Christopoulos GI, O'Doherty JP, Dolan RJ, Schultz W (2009) Risk-dependent reward value signal in human prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America* 106:7185–7190.

Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. *Science* 307:1642–1645.

Tolhurst DJ, Movshon JA, Dean AF (1983) The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research* 23:775–785.

BIBLIOGRAPHY

Tolias AS, Ecker AS, Siapas AG, Hoenselaar A, Keliris GA, Logothetis NK (2007) Recording chronically from the same neurons in awake, behaving primates. *Journal of Neurophysiology* 98:3780–3790.

Treue S, Hol K, Rauber HJ (2000) Seeing multiple directions of motion-physiology and psychophysics. *Nature Neuroscience* 3:270–276.

Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN (2005) A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology* 93:1074–1089.

Tsien RY (1989) Fluorescent probes of cell signaling. *Annual Review of Neuroscience* 12:227–253.

Tsodyks M, Kenet T, Grinvald A, Arieli A (1999) Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science* 286:1943–1946.

Van Zandt T (2000) ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26:582–600.

Ventura V, Cai C, Kass RE (2005) Trial-to-trial variability and its effect on time-varying dependency between two neurons. *Journal of Neurophysiology* 94:2928–2939.

Verhaegen M, Verdult V (2007) *Filtering and system identification: a least squares approach.* Cambridge university press.

BIBLIOGRAPHY

Vickers D (1979) *Decision processes in visual perception.* Academic Press, New York.

Vidne M, Ahmadian Y, Shlens J, Pillow JW, Kulkarni J, Litke AM, Chichilnisky EJ, Simoncelli E, Paninski L (2012) Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *Journal of Computational Neuroscience* 33:97–121.

Vladimirov N, Mu Y, Kawashima T, Bennett DV, Yang CT, Looger LL, Keller PJ, Freeman J, Ahrens MB (2014) Light-sheet functional imaging in fictively behaving zebrafish. *Nature Methods* 11:883–884.

Vogels R, Spileers W, Orban GA (1989) The response variability of striate cortical neurons in the behaving monkey. *Experimental Brain Research* 77:432–436.

Vogelstein JT, Packer AM, Machado TA, Sippy T, Babadi B, Yuste R, Paninski L (2010) Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of Neurophysiology* 104:3691–3704.

Vogelstein JT, Watson BO, Packer AM, Yuste R, Jedynak B, Paninski L (2009) Spike inference from calcium imaging using sequential Monte Carlo methods. *Biophysical Journal* 97:636–655.

Volkmann J (1934) The relation of the time of judgment to the certainty of judgment. *Psychological Bulletin* 31:672–673.

BIBLIOGRAPHY

Wang K, Sun W, Richie CT, Harvey BK, Betzig E, Ji N (2015) Direct wavefront sensing for high-resolution in vivo imaging in scattering tissue. *Nature Communications* 6:7276.

Wang XJ (2002) Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* 36:1–14.

Wang XJ (2008) Decision making in recurrent neuronal circuits. *Neuron* 60:215–234.

Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11:3571–3594.

Wei Z, Wang XJ (2015) Confidence estimation as a stochastic process in a neurodynamical system of decision making. *Journal of Neurophysiology* 114:99–113.

Whiteley L, Sahani M (2008) Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision* 8:2.1–15.

Wilt BA, Burns LD, Ho ETW, Ghosh KK, Mukamel EA, Schnitzer MJ (2009) Advances in light microscopy for neuroscience. *Annual Review of Neuroscience* 32:435.

Wong KF, Huk AC, Shadlen MN, Wang XJ (2007) Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision making. *Frontiers in Computational Neuroscience* 1:6.

BIBLIOGRAPHY

Wong KF, Wang XJ (2006) A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience* 26:1314–1328.

Xu NL, Harnett MT, Williams SR, Huber D, OConnor DH, Svoboda K, Magee JC (2012) Nonlinear dendritic integration of sensory and motor input during an active sensing task. *Nature* 492:247–251.

Yaksi E, Friedrich RW (2006) Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca^{2+} imaging. *Nature Methods* 3:377–383.

Yang T, Shadlen MN (2007) Probabilistic reasoning by neurons. *Nature* 447:1075–1080.

Yang W, Miller JeK, Carrillo-Reid L, Pnevmatikakis E, Paninski L, Yuste R, Peterka DS (2016) Simultaneous multi-plane imaging of neural circuits. *Neuron* 89:269–284.

Yasuda R, Nimchinsky EA, Scheuss V, Pologruto TA, Oertner TG, Sabatini BL, Svoboda K (2004) Imaging calcium concentration dynamics in small neuronal compartments. *Science Signaling* 2004:1–20.

Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M (2009) Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology* 102:614–635.

Yu J, Gutnisky DA, Hires SA, Svoboda K (2016) Layer 4 fast-spiking interneu-

rons filter thalamocortical signals during active somatosensation. *Nature Neuroscience* 19:1647–1657.

Yuste R, Majewska A, Cash SS, Denk W (1999) Mechanisms of calcium influx into hippocampal spines: heterogeneity among spines, coincidence detection by nmda receptors, and optical quantal analysis. *Journal of Neuroscience* 19:1976–1987.

Zhao Y, Park IM (2016) Variational latent gaussian process for recovering single-trial dynamics from population spike trains. *arXiv:1604.03053* .

Zhou K, Doyle JC, Glover K et al. (1996) *Robust and optimal control*. Prentice Hall New Jersey.

Zylberberg J, Murphy JT, DeWeese MR (2011) A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields. *PLoS Computational Biology* 7:e1002250.

The Johns Hopkins University School of Medicine

Ziqiang Wei Jan. 27, 17

Education:

Ph.D. expected	2017	Program in Neuroscience	Johns Hopkins School of Medicine
		Mentor: Shaul Druckmann PhD (Janelia Research Campus, HHMI)	
M.S.	2010	Control Theory	Chinese Academy of Sciences
B.S.	2007	Management Science	Beijing Normal University

Research rotation	2011-2014	Lab of Dmitri Chklovskii, Janelia Research Campus
Research rotation	2011-2011	Lab of Ernst Niebur, Johns Hopkins School of Medicine
Research rotation	2010-2011	Lab of Rudiger von der Heydt, Johns Hopkins School of Medicine

Publications:

Wei Z and Wang XJ. (2015). Confidence estimation as a stochastic process in a neurodynamical system of decision making. J. Neurophys., 114, 99-113.

Wei Z, Wang XJ, and Wang DH. (2012). From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. J. Neurosci., 32, 11228-11240.

Wei Z, Hong Y, and Wang D. (2009). The phase diagram and the pathway of phase transitions for traffic flow in a circular one-lane roadway. Physica A, 388, 1665-1672.

Wang D, **Wei Z**, and Fan Y. (2007). Hysteresis phenomena of the intelligent driver model for traffic flow. Phys. Rev. E, 76, 016105.

Posters:

Wei Z, Wan Y, Keller P, and Druckmann S. (2016). Analyzing ensemble activity in the developing zebrafish spinal cord. Society for Neuroscience Abstract, 215.10, San Diego, CA.

Wan Y, **Wei Z**, Druckmann S, and Keller P. (2016). Emergence of patterned activity in the developing zebrafish spinal cord. Society for Neuroscience Abstract, 215.21, San Diego, CA.

Wei Z, Li N, Svoboda K, and Druckmann S. (2014). Neural correlates of motor planning in a delayed response task. Society for Neuroscience Abstract, 735.01, Washington, DC.

Wei Z, Hu T, and Chklovskii DB. (2013). Optimal de-noising and predictive coding account for spatiotemporal receptive field of LGN neurons. Computational and Systems Neuroscience Meeting, III-78, Salt Lake City, UT.

Wei Z, Hu T, and Chklovskii DB. (2012). Efficient coding of natural scenes in the early visual system. Society for Neuroscience Abstract, 631.11, New Orleans, LA.

Wei Z, Hu T, and Chklovskii DB. (2012). Biologically plausible learning of sparse-coding dictionary in a neural network. Computational and Systems Neuroscience Meeting, I-23, Salt Lake City, UT.

Wei Z, Wang D, and Wang XJ. (2011). From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. Society for Neuroscience Abstract, 811.03, Washington, DC.

Wei Z, Mihalas S, and Niebur E. (2011). Representation of temporal information in a neural circuitry model of working memory. Society for Neuroscience Abstract, 811.17, Washington, DC.

Wei Z, Hu T, and Chklovskii DB. (2011). Emergence of sparse representation in a neural network through Hebbian Learning. Sloan-Swartz Annual Meeting, Ashburn, VA.

Wei Z, and Wang XJ. (2010). Choices under uncertainty: Confidence representation in a neural network of decision making. Society for Neuroscience Abstract, 631.11, San Diego, CA.